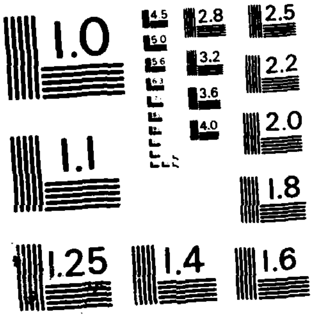
AD-A131 048 LIMITED CONNECTED SPEECH EXPERIMENT(U) ITT DEFENSE COMMUNICATIONS DIV NUTLEY N J P B LANDELL MAR 83 RADC-TR-82-316 F30602-81-C-0155 1/2

UNCLASSIFIED F/G 17/2 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

12

[illegible]
Final Technical Report
March 1983

LIMITED CONNECTED SPEECH EXPERIMENT

BY

Patrick [illegible]

[illegible]

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-82-316 has been reviewed and is approved for publication.

APPROVED: John V. Ferrante
JOHN V. FERRANTE
Project Engineer

APPROVED:
JOHN N. ENTZMINGER, JR.
Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER:
[illegible]
[illegible]

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-82-316	2. GOVT ACCESSION NO. AD-A131 048	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) LIMITED CONNECTED SPEECH EXPERIMENT		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report 28 May 81 - 20 Aug 81
		6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) Patrick B. Landell		8. CONTRACT OR GRANT NUMBER(s) F30602-81-C-0155
9. PERFORMING ORGANIZATION NAME AND ADDRESS ITT Defense Communications Division 492 River Road Nutley NJ 07110		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62702F 45941581
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRAA) Griffiss AFB NY 13441		12. REPORT DATE March 1983
		13. NUMBER OF PAGES 114
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) Same		
18. SUPPLEMENTARY NOTES RADC Project Engineer: John V. Farrante, Capt, USAF (IRAA)		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Connected Word Recognition Template Extraction Template Averaging Syntax Control	Data Base Design	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This document is the final report from ITTDCD to RADC for Contract Number F30602-81-C-0155 entitled "Limited Connected Speech Experiment, (LCSE)". The purpose of this contract was to demonstrate that Connected Speech Recognition (CSR) can be performed in real-time on a vocabulary of one hundred words and to test the performance of the CSR system for twenty-five male and twenty-five female speakers. This report describes ~~ITTDCD~~'s real-time laboratory CSR system, the data base and training software de-

veloped in accordance with the contract, and the results of the performance tests.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. EXECUTIVE SUMMARY

This document is the final report from ITTDCD to RADC for Contract Number F30602-81-C-0155 entitled Limited Connected Speech Experiment(LCSE). The purpose of this contract was to demonstrate that Connected Speech Recognition (CSR) can be performed in real time on a vocabulary of one hundred words and to test the performance of the CSR system for twenty five male and twenty five female speakers. This report describes ITTDCD's real time laboratory CSR system, the data base and training software developed in accordance with the contract, and the results of the performance tests.

ITTDCD's real time laboratory system is a flexible speech recognition program which operates in an FPS AP-120B array processor that is connected to a VAX 11/780 computer. The user can easily define the vocabulary and syntax for a given recognition task via interactive syntax specification commands. In addition to performing task specific phrase recognition, the CSR program has a "voice-control" feature which allows the user to control the system via spoken commands of his own choosing. A versatile training capability permits the user to adapt to the speaker dependent system by speaking both words and phrases from a vocabulary which he has defined. The CSR system is also a valuable research and development tool with analysis mode and recognition experiment mode features.

An airline query task was chosen to define the 100 word recognition vocabulary for the LCSE data base. The phrases associated with this task are representative of a simplified air travel information retrieval application. This syntax and vocabulary were designed primarily with the goal of user flexibilty in the task and not with the goal of optimal recognition performance. The vocabulary includes three phonetically similar groups of words: the digits, the teens ("ten" through "nineteen"), and the decades ("twenty" through "ninety"). The vocabulary also includes three function words "of", for", and "the" which are often unstressed in continuous speech. Analog recordings were made for 25 males and 25 females, each speaking words and phrases from the 100 word vocabulary.

Template training is a critical step for speaker dependent CSR systems and a major accomplishment of the Limited Connected Speech Experiment was the development of two effective training techniques, template extraction and template averaging. ITTDCD's template extraction algorithm automatically locates and saves the speech parameters of words embedded in continuous phrases. The template averaging technique performs a clustering analysis on multiple tokens for the same word and averages the speech parameters of similar tokens. Training tokens output by the template extraction process are input to the template averager along with tokens of individual words spoken in isolation. Each of the 50 data base subjects spoke three repetitions of the 100 word vocabulary as well as 66 phrases which could be used for template extraction.

The performance tests were conducted for fifty speakers each speaking 50 random phrases from the airline query grammar. The phrases contained 7.4 words on average. The median word recognition accuracy was 94.5% for all words. Ignoring errors on the words "of","for", and "the", the median word rate was 96.8% and the median phrase recognition rate was 84%. A phrase is considered correct if all words are correcty identified. An extensive error analysis of the performance test results was undertaken with all word errors being assigned to ten error classification types.

Function words aside, the major cause of word errors was found to be confusions amongst the digits, decades, and teens. Typical examples are confusions between "seven" and "seventy", "eight" and "eighty", and "sixty" and "sixteen". In order to examine the impact of vocabulary selection on recognition performance for a given task, another performance test was designed and conducted with an 82 word version of the airline query grammar. The decade and teen nodes were eliminated from the syntax and all test phrases containing decade or teen words were eliminated, reducing the average number of test utterances per speaker from 50 to 32. Excluding "of", for", and "the" errors, a 98.0% word rate and 90.5% phrase rate was achieved on the 82 word vocabulary test.

On August 20, 1982, a demonstration of ITTDCD's real time laboratory CSR system was presented to a representative of RADC. Six of the fifty performance test subjects were asked to speak a series of phrases from the 100 word airline query grammar. These phrases were recognized with an accuracy comparable to that achieved on the performance test. In the course of the demonstration, various features of the CSR system were exhibited including voice-control, training, template extraction, template averaging, and analysis.

2. INTRODUCTION

The Limited Connected Speech Experiment had two primary goals. First, to demonstrate a Connected Speech Recognition (CSR) system which provides real time response on a vocabulary of 100 words. And secondly, to test the recognition performance of this CSR system over a data base consisting of 25 male and 25 female speakers. This chapter gives an overview of the tasks that were carried out to achieve these goals

2.1 Development of CSR Control Software

Executive software was developed on the VAX computer to control the overall operation of the CSR system including training, recognition, experiment, and analysis. In addition software was developed to provide for creation and maintenance of speech parameter files for word templates and phrases. A description of the operation of the CSR system appears in Chapter 3 along with a brief description of its recognition algorithm.

2.2 Development of Syntax Specification Software

An interactive syntax specification program was designed and implemented. This software enables the operator to specify the vocabulary words, and the finite state grammar which define a CSR task. The resulting syntax file is employed to guide training and recognition software on the CSR system. Further detail on syntax specification is presented in Chapter 3.

2.3 Selection of a Syntax and Vocabulary

An airline query task was chosen to define the 100 word recognition vocabulary for the LCSE data base. It was designed to be representative of limited syntax application areas for speech recognition. The phrases are representative of a simplified air travel information **retrieval** application. The entire 100 words and associated finite state syntax are presented in detail in Chapter 4 of this report.

2.4 Generation of the Data Base.

Analog recordings were made for 25 males and 25 females, each speaking words and phrases from the 100 word airline query grammar. About half of the speakers were chosen from within ITTDCD's San Diego laboratory and the remainder were selected from agency referrals. None of the speakers had any prior experience with speech recognition systems. Files of digital speech parameters for each word and phrase were obtained by playing the analog tapes into a filterbank. The steps taken to generate this data base are discussed in Chapter 4.

2.5 Investigation of Template Averaging Techniques

After a review of the literature, a template averaging technique was implemented and tested on an existing data base of connected digit phrases from five speakers. To obtain a performance baseline for evaluating the technique, recognition experiments were performed using single tokens as templates for each word. CSR experiments were then conducted with averaged templates. A description of the template averaging technique is contained in Chapter 5 along with results of the template averaging study.

2.6 Extraction Templates

Software was developed for automatically extracting templates from connected speech utterances for the purpose of combining them with existing templates of the same vocabulary word. The recognition system itself controls the template extraction process, as described in Chapter 5.

2.7 Integration and Test of Software

Training and recognition software were integrated on the VAX - AP120B system. Ten of the data base subjects were designated as development speakers and a series of recognition experiments were performed to establish the appropriate training technique for the 50 speaker performance test. Chapter 6 describes the experimental findings of the development testing process.

2.8 Performance Test

Templates were prepared for the 50 speaker data base using the template extraction and template averaging software. 50 phrases from the airline query grammar were input to the CSR system for each of the 50 speakers. Performance test results are presented in Chapter 7 and an analysis of word recognition errors is addressed in Chapter 8.

2.9 Demonstration

A demonstration of the CSR system was prepared and conducted for government representatives. Six speakers from the performance test group participated in the demonstration.

3. A REAL TIME LABORATORY CONNECTED SPEECH RECOGNITION SYSTEM

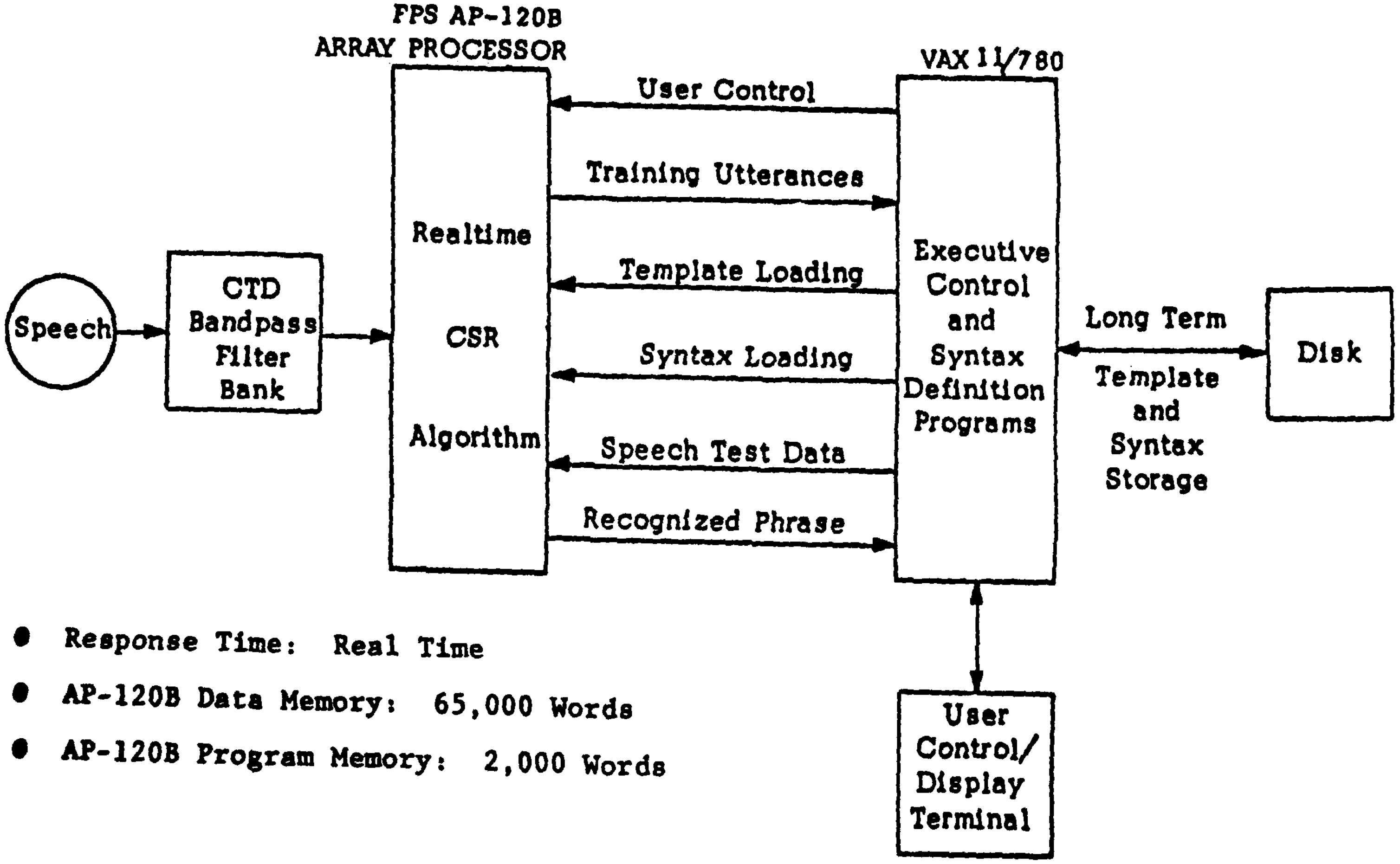
The Connected Speech Recognition (CSR) system was developed at ITTDCD's San Diego laboratory to accomplish the goal of recognizing a syntax constrained, 100 word vocabulary in real time with a low error rate. This chapter gives an overview of the operation and features of the system and a brief description of the CSR recognition algorithm upon which the system is based.

3.1 Operational Overview*

Figure 3.1 shows a diagram of the limited connected speech exploratory development system. The CSR system operates in an FPS AP-120B array processor which is connected to a VAX 11/780 computer. The recognition algorithm is contained entirely in the array processor with the VAX serving as the executive controller which handles the user interface, long term storage of templates and grammars, and support software such as the template averaging routines. An analog filterbank is connected to the array processor via a DMA channel to permit realtime processing of the speech signal.

The user controls the system by keyboard input at a display terminal and by a set of single word voice commands. At any time, the system is in one of three distinct states, as illustrated in the state diagram of Figure 3.2. These states are called the Command state, the Recognition state, and the Voice Control state. The Command state is the normal, or default state of the system. In this state, 25 different commands can be entered from the keyboard or read from specified command files. Some of these commands do the following:

*A complete users guide to the CSR system is included as Appendix A.

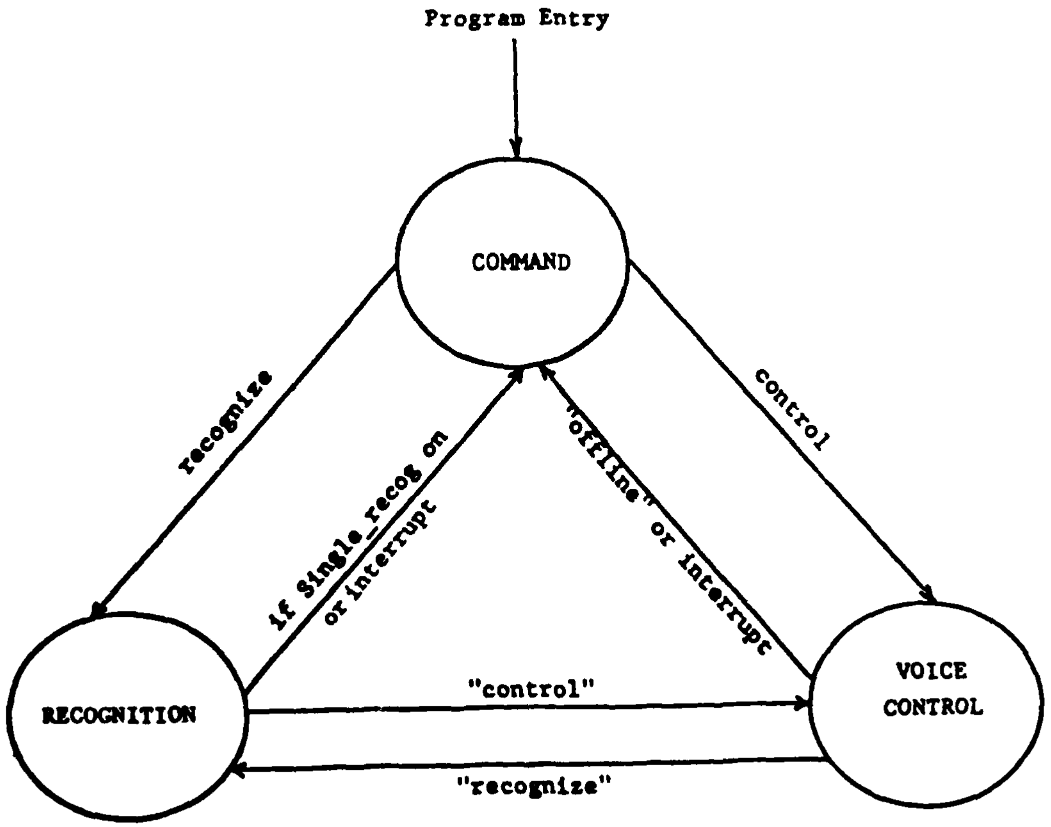


- Response Time: Real Time
- AP-120B Data Memory: 65,000 Words
- AP-120B Program Memory: 2,000 Words

FIGURE 3-1
LIMITED CONNECTED SPEECH EXPLORATORY DEVELOPMENT MODEL

FIGURE 3-2

CSR System State Diagram



- Read the syntax from a specified file,
- Set an environment variable (eg.,vocabulary size),
- Train the vocabulary, or a particular word,
- Turn the experiment mode on and collect statistics,
- Execute a specified command file (which contains commands like these),
- Execute a VAX 11/780 operator command and return to the CSR system,
- Exit from the CSR system.

In addition to these commands, executing a recognize command changes the state of the system to the Recognition state, and executing the control command changes it to the Voice Control state.

In the Recognition state, the system will recognize any syntactically legal phrase specified by the grammar and the vocabulary of the current task. After the phrase is spoken the recognized text is displayed on the terminal and the system either returns to the Command state, if the environment variable "Single_recog" has been set on, or remains in the Recognition state ready for the next utterance, if the variable has been set off. In the latter mode the user may return to the command state by hitting the interrupt key. At any point, while speaking a phrase, the user may cancel the phrase by immediately saying "Cancel". A transition to the Voice Control state is accomplished by saying the word "Control".

In the Voice Control state a small subset of the 25 commands available to the user in the Command state are activated by voice. When the user trains the system, he is prompted to speak these control words, along with the task dependent vocabulary words. The voice commands currently implemented are:

- "Display": Displays the current values of the environment variables.
- "Options": Displays the five best scoring phrase recognition options for the last task phrase spoken in the Recognition state.
- "Word-scores": Displays the individual word scores for the last task phrase spoken in the Recognition state.
- "Recognize": Change the system state to the recognition state.
- "Offline": Release the array processor and return to the command state.

As in the Recognition state, the user can also switch from the Voice Control state to the Command state by hitting the interrupt key.

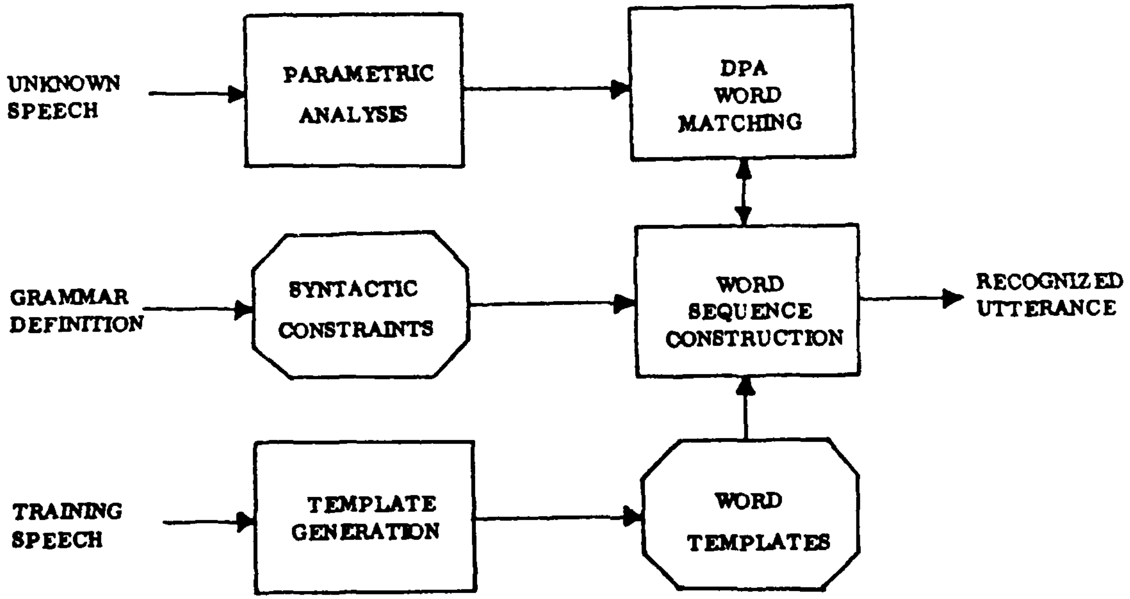
3.2 The ITTDCD CSR Algorithm

Figure 3-3 gives an overview of the ITTDCD CSR algorithm. Three types of inputs are supplied to the system, as shown on the left-hand side of the figure. The input speech undergoes a parametric analysis performed by a Charge Transfer Device (CTD) band pass filterbank. This filterbank was previously developed in conjunction with an earlier RADC contract, the Solid State Audio/Speech Processor Analysis (SSA/SPA) Contract (No. F30602-78-C0359). Using eighteen 1/3 octave switched-capacitor band pass filters and one full octave filter the filterbank covers a frequency range of 100Hz to 9500 Hz and supplies 19 coefficients every 10 ms to a parameter reduction algorithm.

The parameter reduction algorithm performs variable frame rate encoding to remove redundant frames and converts the parameters to ten mel-cepstral coefficients using a mel-cosine linear transformation. Details of this algorithm can be found in sections 2.1.2.2 and 3.2.3 of the SSA/SPA final report.

FIGURE 3-3

AN INITIAL CONTINUOUS SPEECH RECOGNITION SYSTEM



The coefficients from the parametric analysis step are passed to the word matching algorithm. This algorithm compares each word template with the spoken utterance using a non-linear time alignment process carried out by the dynamic programming match algorithm. The non-linear time alignment is necessary to account for the natural time variations between different utterances of the same word. The time warp constraints used in the algorithm force the length of the spoken word to be between one-half and twice the length of its template.

A second level dynamic programming algorithm is implemented in the Word Sequence Control block to control word template matching and to concatenate matched templates into the connected word sequence which best matches the input utterance. Syntactic constraints define the set of word sequences that can be recognized by the system as sentences.

The ITTDCD CSR algorithm processes an unknown utterance from left to right to find a sequence of words that closely matches it. Disregarding syntactic control for now, the process takes place as follows: The dynamic programming algorithm processes the unknown utterance one frame at a time. At every frame, matching begins for all word templates. After a delay of ½ of the length of a word template each frame is a possible ending point for the template started. Thus, at each frame, F, in which the matching of some word template, W, ends, a set of candidate partial phrases (word sequences) is formed by appending the word W to the set of partial phrases ending where W began. This is done for every word ending at frame F, resulting in a large set of candidate partial phrases ending at the frame. Because of memory and processing limitations, only the best N candidates are retained, where N is determined for each frame based on the scores of the competing candidates. This technique of varying the number of candidate phrases considered at each frame is called a beam search [Lowerre and Reddy - 1980]. With the beam search strategy, the system allocates more of its resources to nodes in the

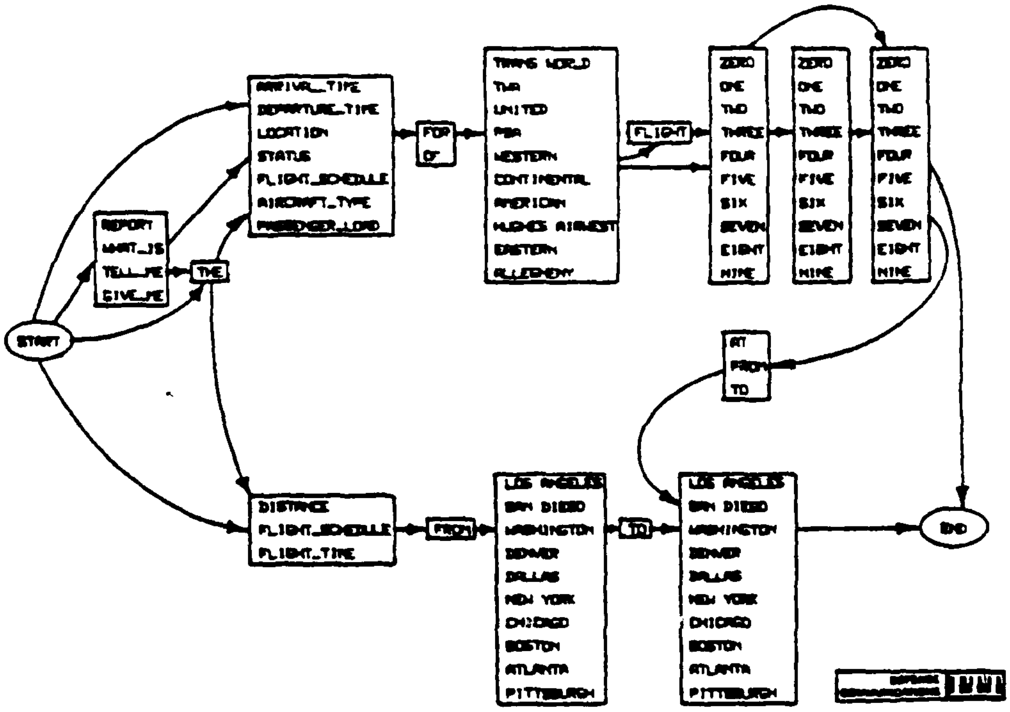


Figure 3-4
An Example of a Finite State Grammar

node-to-node connection matrix which completely describe the syntax.

Figure 3-5 expands in detail the operation of the Word Sequence Control Module appearing as a simple block in the algorithm overview of Figure 3-3. The control module keeps track of the best candidate phrases ending at each frame. At every frame of the input utterance, the competing partial phrases are stored in the phrase description tables shown near the bottom of the figure. Words matching the previous portion of the input utterance are used to extend partial phrases from the table to obtain a new set of partial phrases. The new set is limited by the beam search strategy and stored in the phrase description tables. The grammar node states specified by the candidate phrases thus determine which nodes and words will be processed in the next frame. The new grammar node states are expanded into a set of new template candidates by using word-node membership information. These new template candidates will be used by the recognition algorithm to match the next part of the input utterance.

When the end of the utterance is detected, the next grammar node states are checked to determine which of them are connected to a final state. The best scoring candidate phrase leading to a final state is reported as the recognized sentence.

The templates used to match the input utterance are obtained from the template generation software from training speech. The techniques used in this part of the system are the discussed in Chapter 5.

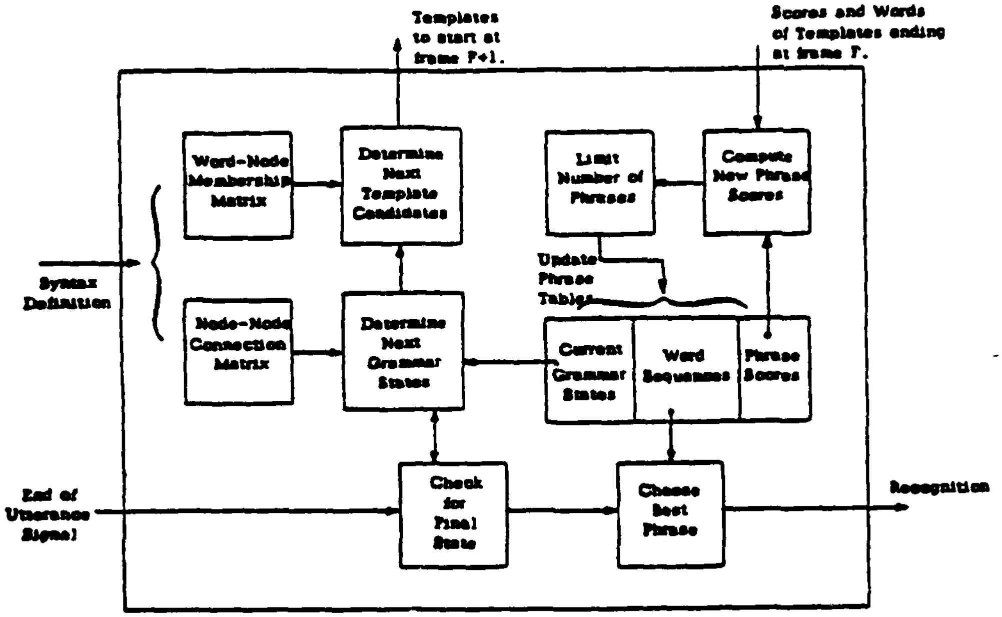

Templates to start at frame F+1.
Scores and Words of Templates ending at frame F.
Word-Node Membership Matrix
Determine Next Template Candidates
Limit Number of Phrases
Compute New Phrase Scores
Syntax Definition
Update Phrase Tables
Node-Node Connection Matrix
Determine Next Grammar States
Current Grammar States
Word Sequences
Phrase Scores
End of Utterance Signal
Check for Final State
Choose Best Phrase
Recognition

Figure 3-5
Details of Word Syntactic Analysis

4. THE LIMITED CONNECTED SPEECH DATA BASE

The development of a connected speech recognition system requires extensive testing on large data bases from a wide sampling of the speaker population to obtain statistically significant performance data. For the LCSE work ITTDCD recorded three repetitions of a 100 word vocabulary and 166 sentences generated from the vocabulary for each of 25 male and 25 female speakers. Generation of this data base required seven carefully performed tasks. They are:

1. Data base design,
2. Design and acquisition of recording facilities,
3. Design and implementation of data base collection software,
4. Selection of a speaker population,
5. Data base collection,
6. Data base pruning,
7. Data base processing.

These seven tasks will be discussed in this chapter and the related Append.x B.

4.1 LCSE Data Base Design

In response to the contract requirements a limited syntax 100 word vocabulary data base was designed to be recorded by 25 females and 25 males. Figure 4-1a shows the finite state syntax node structure and Figure 4-1b gives the vocabulary words assigned to each node. The syntax and vocabulary was chosen to be representative of a simplified air travel information retrieval application which we call the airline ɥuery task. However, the syntax and vocabulary were designed pri-

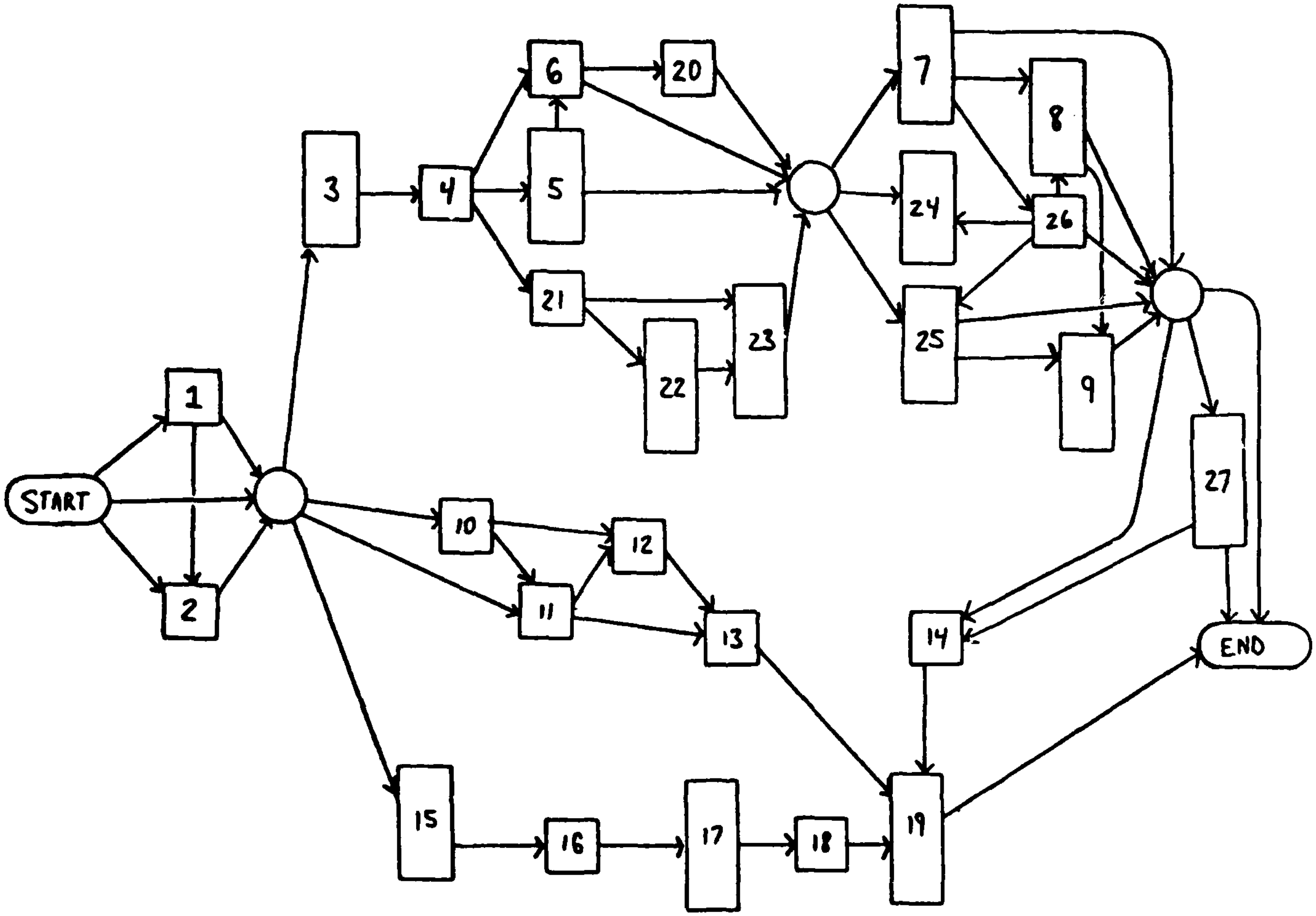


Figure 4-1a
Syntax Node Structure 100 Word Airline Query Grammar

Figure 4-1b

Assignment of the 100-Word Vocabulary to Nodes

1. Report
What-is
Tell-me
Give-me

2. The

3. Arrival-time
Departure-time
Location
Status
Flight-Schedule
Aircraft-Type
Passenger-Load

4. For
Of

5. National
TWA
United
PSA
Western
Continental
American
Hughes-Airwest
Eastern
Allegheny

6. Flight

7. Zero
One
Two
Three
Four
Five
Six
Seven
Eight
Nine

8. Same as Node 7.

9. Same as Node 7.

10. Current

11. Weather

12. Forecast

13. For
Of

14. At
From
To

15. Distance
Flight-Schedule
Flight-Time

16. From

17. Los Angeles
San Diego
Washington
Denver
Dallas
New York
Chicago
Boston
Atlanta
Pittsburgh

18. To

19. Same as Node 17.

20. Number

21. Aircraft

22. Alpha
Bravo
Charlie
Delta
Echo
Foxtrot
Golf
Hotel
India
Juliet
Kilo
Lima
Mike
November
Oscar
Papa
Quebec
Romeo
Sierra
Tango
Uniform
Victor
Whiskey
X-Ray
Yankee
Zulu

23. Same as Node 22.

24. Ten
Eleven
Twelve
Thirteen
Fourteen
Fifteen
Sixteen
Seventeen
Eighteen
Nineteen

25. Twenty
Thirty
Forty
Fifty
Sixty
Seventy
Eighty
Ninety

Hundred

27. Same as Node 22

marily with the goal of user flexibilty in the task and not with the goal of optimal recognition performance. Thus, the vocabulary includes the digits, the teens ("ten" through "nineteen"), and the decades ("twenty" through "ninety"), which can be spoken in various combinations within an airline query phrase. The similarity of digit, teen, and decade words can pose a challenging recognition task since the syntax allows any of these words to appear in the same place in a phrase.

The 100 word vocabulary also includes 26 alpha words ("alpha", "bravo", "charlie", etc.), which comprise one node. This node both precedes and follows the digit-teen-decade sequence in the syntax and provides a test of the systems ability to match many templates against the input utterance in real time.

The LCSE connected data base was designed to be subpart of a larger data base collection effort. The larger data base included a 200 and 300 word limited syntax airline query grammar, a connected digit component, an alphabet spelling component, and a diagnostic rhyme component. The content of this data base, the recording facilities, and the procedures use to collect the data have been described in a paper* given at the Workshop on Standardization for Speech I/O Technology on March 18, 1982. This paper covers the first five topics listed above for the LCSE connected speech data base and is included in Appendix B. The last two topics of data base pruning and data base processing are discussed in the remainder of this chapter.

4.2 Data Base Pruning

A total of 63 speakers (31 males and 32 females) were recorded to permit selection of speech data for 25 males and 25 females which is free from recording and processing problems. After recording, the data base was pruned from 63

*Landell,B. P., Smith, A. R., Koble, H. M., and Alcove, M. L., "A Continuous Speech Data Base," presented at the Workshop on Standardization for Speech I/O Technology, National Bureau of Standards, Gaithersburg, Md., March 18-19, 1982.

speakers to 50 speakers. This pruning was done before any recognition performance testing and was based on either random elimination, or on the presence of difficulties in the recording procedures. The following table lists the reasons for speaker elimination due to recording difficulties:

- For the first speaker (male) the recording procedures had not been completely tested so that the recording session was very long and fragmented,
- One speaker (female) did not complete the recording session,
- Excessive tampering with the close talking head mounted microphone during the session (2 males),
- Tape recorder was set with the variable pitch control activated (2 males),
- Part of analog tape was recorded over (1 male),
- Excessive environment noise outside of recording room (3 females).

Although some of these problems may in fact be conditions which a speech system might encounter (eg., environmental noise and microphone movement), they were not conditions that we wanted to study in this contract. After pruning speakers with recording difficulties, the resultant data base contained 25 males and 28 females. The remaining three females were eliminated by random selection.

4.3 Data Base Processing

The above data base generation steps resulted in a set of analog tapes containing training and test data for 50 speakers. Although these tapes could have been used directly in the system to train it for each speaker and then to test it, such a procedure would require many hours of tape handling as various tests were run. These problems were circumvented by processing the data once through the filterbank to create a series of digital files containing word training tokens and test phrases. These files were then used over and over again by computer programs to train and test the system as it was developed.

The data base processing task is a two step process. In the first step the operator plays about five minutes of speech (determined by available disk space) from an audio tape through the filterbank/array processor front end to generate an output file of filter parameters. In the second step, a VAX 11/780 program processes the filter parameter file using the recording session history file*. The VAX program uses marking tones in the filter parameter file to synchronize the frame position with time marks in the history file. The VAX program also performs endpoint detection to find each utterance in the output file and splits the input filter parameter file into smaller files which are tagged with an ASCII label describing the utterance. Occasional endpoint detection problems occured when a speaker corrected himself in midst of a word or phrase without pausing before repeating the text. In these cases (less than 1% of all utterances), the operator listened to the speech and used an amplitude plot to determine the proper window within which the endpoint detection algorithm could be safely rerun.

*As explained in Appendix B, the recording history file contains the exact text with which the speaker was prompted together with a time mark. The time mark is computed relative to a tone recorded on the tape at the beginning of the session.

5. TRAINING TECHNIQUES

Template training is a critical step for speaker dependent CSR systems, since the performance of the system is limited by the degree to which the templates model the test speech. This chapter describes two techniques that were used to obtain improved word template models. They are template averaging and the extraction of templates from connected speech.

5.1 Template Averaging Study

A template averaging study was performed to evaluate the effectiveness of template clustering and averaging techniques in connected speech recognition. After a review of the literature, we decided to employ the Unsupervised Clustering Without Averaging (UWA) algorithm as described by Rabiner and Wilpon* Since complete details of the technique are available in their paper we will only give an overview of the technique before discussing our results.

The technique was implemented on the PDP 11/60 to cluster and average multiple tokens or samples for a given word. Inputs to this software include the file names of individual tokens, a clustering distance threshold, and the number of desired output templates. The software performs three steps to obtain the averaged templates: first, it computes the similarity distance and the frame-to-frame correspondence between each pair of tokens, second, it applies a clustering algorithm to the tokens, and finally, it averages the speech parameters across the tokens of each cluster on a frame-by-frame basis according to the previously

*Rabiner, L., and Wilpon, J., "Considerations in Applying Clustering Techniques to Speaker-Independent Word Recognition," Journal of the Acoustical Society of America, 66 (3), September 1979.

computed frame-to-frame correspondence.

In the first step a dynamic programming algorithm (DPA) is used to compute the similarity distance and the frame-to-frame correspondence between pairs of input tokens. The DPA finds the best non-linear correspondence between a pair of tokens. The average distance between the frames that are aligned by the process gives the overall similarity distance between the tokens. Constraints are imposed on this alignment process so that either token cannot be "stretched" more than twice its length to match the other token. Thus, in some cases tokens do not match and are given a large similarity distance. In the clustering step which follows, these tokens will be prevented from appearing in the same cluster. The algorithm thus prevents such tokens from being averaged.

The similarity distances between all token pairs give an intertoken distance matrix. The cluster step of the algorithm uses the matrix to compute the minimax center of the token set. The minimax center is simply that token for which the maximum distance to any other token is minimized. Then, in an iterative process, any token whose distance to the center exceeds the clustering threshold (supplied to the algorithm) is removed. A new minimax center is then computed on the reduced set. The process iterates until the center does not change. All tokens within the final set are within the cluster threshold and form the cluster and the final center token becomes cluster center. Tokens that have been removed are reprocessed to find a second cluster. The process continues until no tokens remain. A variable number of clusters are computed and each token is finally assigned to a cluster (an outlying token might form a cluster of one).

The final averaging step averages all tokens within each cluster. The frame-to-frame correspondence of each token with the center token determines which frames are averaged together to form the final average template.

5.1.1 Preliminary Averaging Experiments with Filterbank Parameters.

Averaged templates were evaluated in several speech recognition experiments. A connected digit data base of five speakers (three male and two female) was used in these experiments. The training data base consisted of six tokens per word per speaker for each of the ten digits. The test data base contained 150 three, four, and five digit phrases per speaker. To provide a baseline for these experiments, each of the six token sets for each speaker were used as templates. The first line in Table 5-1 is the average recognition rate in the baseline experiments. The second line shows the performance using the center of all tokens for each word. That is, the cluster threshold was set at a maximum value so that only one cluster was found. For the third experiment, this threshold was dropped so that more than one cluster may have been formed per word. The center token of the largest cluster was used to represent the word. For the experiment shown on the fourth line, the tokens of the largest cluster were averaged. This experiment showed that a higher phrase recognition rate (76%) could be achieved by averaging parameters than by techniques of selecting individual tokens to represent the words.

Table 5-1
Clustering and Averaging Results
With Filter Bank Parameters

TEMPLATE SET	THRESHOLD	COMBINED PHRASE RATE ALL SPEAKERS	COMBINED WORD RATE ALL SPEAKERS
MEAN RATE OF SIX TOKEN SETS	N/A	67	89
TOKEN CENTERS	MAX	65	88
LARGEST CLUSTER CENTERS	27	70	91
AVERAGED LARGEST CLUSTER	27	76	93

5.1.2 Experiments Averaging Speech Parameters.

As described in Chapter 3, the CSR system uses mel-cepstral coefficients for its speech parameters. These are obtained from the filterbank parameters by a linear transformation During the template averaging study another type of linear transformation which was then under investigation was employed. This transformation was obtained by performing a linear discriminant analysis on marked and labeled speech segments. Although the linear discriminant transformation technique was abandoned (it was too speaker dependent for the Limited CSR system), the results of the averaging study using these parameters are presented here because the parameters are similar to the mel-cepstral coefficients used in the system. Table 5.2 presents results of the experiments. The right hand column labelled "Not Avgd" represents the results of running each token set of linear discriminant parameters separately and averaging the results. This column represents a benchmark for evaluating the effectiveness of the averaging process. The columns headed by the clustering thresholds present the template averaging results. In each case, the tokens in the largest cluster were averaged producing one averaged output template per word and in each case, improved recognition results were obtained in comparison to the *baseline figures*. An additional experiment was performed in which templates made from averaged filter bank parameters (line 4 of Table 5-1) were transformed prior to recognition. This experiment yielded a phrase rate of 90% and a word rate of 97% indicating that comparable recognition results are achieved by averaging transformed parameters than by averaging filter parameters and then applying the linear transformation.

Table 5-2
Experiments Averaging Linear Discriminant Parameters

Cluster	Phrase Recognition Rate (150 Digit Strings per Speaker)						
Threshold → Speaker	23	28	33	44	55	Max	Not Avgd
BB(M)	92	96	96	96	97	97	89
CL(F)	83	91	90	89	88	88	77
DM(F)	93	98	97	96	96	96	87
MM(M)	70	77	85	84	85	85	75
RS(M)	96	96	96	96	94	94	89
Trials →	750	750	750	750	750	750	750
Phrase Rate	87	92	93	92	92	92	84
Word Rate	97	98	98	98	98	98	95

5.1.3 Conclusions

The preliminary experiments indicated that simply using cluster centers as templates did not significantly improve recognition. However, averaging proved to be effective. Averaging of filter bank parameters cut phrase recognition errors by 30% (performance went from 67% to 76%), and averaging the speech parameters were even more effective in that the phrase errors were cut in half (performance went from 84% to 92%).

An unexpected result was the phrase rate of 92% achieved with a maximum clustering threshold. The maximum clustering threshold forced the software to average all six of the input tokens for each word with the exception of those infrequent cases where a test token was over twice or less than half the length of the center token. In this experiment, for many of the words, quite dissimilar tokens were averaged together yet recognition performance did not suffer. This result as well as the general insensitivity of the process to the clustering threshold is probably due to the limited number of tokens per word.

5.2 Template Extraction from Continuous Speech

A word template produced from an isolated pronunciation of a word is often an inadequate representation of how that word appears in continuous speech. However, isolated word training has the advantage of letting a new speaker quickly and easily train the system. Therefore, our approach allows a speaker to initially train the system by reciting the word vocabulary (in response to system prompting) and then train the system using sentences from the task grammar. Thus, both isolated and continuous versions of a word can be obtained to generate more robust templates.

Associated with a given syntax is a set of standard phrases which are constructed to both satisfy the node sequence of the grammar and to contain one or more occurrences of each word in the vocabulary. During training, the user is prompted to speak each of the phrases in the standard phrase set. The user may also construct and say phrases of his own choosing during the training or retraining phase. When the extract command is executed, the system performs what is called "forced recognition", for each phrase which has been spoken.

"Forced recognition" limits the CSR system so that it can only recognize the sequence of words spoken in the phrase. This is easily obtained by automatically devising a syntax that allows only one sequence of words, i.e., the words of the phrase that has been spoken. The algorithm requires that an existing template be available for each word in the known phrase. Forced recognition is done in the digital input mode, that is, the speech parameters input to the CSR system are read from a phrase file which was created during training. Following the forced recognition, the word endpoints found by the DPA matching module of the recognition algorithm are then used to extract word patterns from the parametric representation of the phrase and these patterns are output as word templates. Before extracting the parameters for a word, the system checks the word scores of the bounding words in the phrase to insure that they are below a threshold. This boundary test insures proper alignment for the extracted word. If the test fails, the word is rejected and not used in the template averaging process.

6. DEVELOPMENT TESTING

The goal of development testing was to identify areas of recognition algorithm improvement by exercising the CSR system and to determine an adequate training technique for the performance test. The intention was that at the conclusion of development testing, the CSR algorithm and the training technique would be established.

6.1 Development Data Base

Ten of the 50 data base subjects were selected as development testing speakers, five males and five females. Each of these speakers had recorded 100 phrases from the 100 word airline query grammar. These 100 phrases were divided into two sets, one to be used as development test material, and a second set to be used in the final performance test. For training material, each speaker had recorded three repetitions of the 100 word vocabulary (these are referred to as "isolated" tokens) and 66 standard phrases available for template extraction. Development experiments were structured so that the 50 development set phrases were run versus one set of templates for each of the ten speaker. However, for six of the speakers one phrase was eliminated because it was syntactically incorrect. Each development experiment thus consisted of 494 phrase trials of average length 7.3 words.

6.2 Function Words.

Six words of the 100 word airline vocabulary are referred to as function words. They are the article "the" and the prepositions "of", "for", "to", "at", and "from".

These words have a special role in the vocabulary and were the target of specific training techniques. The function words are often discussed as a group in this and the remaining chapters cf this report.

From previous experience in continuous speech recognition, we realized that the function words were deserving of special attention. These words are often unstressed and sometimes dropped completely from spoken phrases. Function words are also often significantly colored by coarticulation. Isolated renditions of these words are of little value as templates because their duration is often two to three times longer than function word duration in continuous speech. Thus template extraction seemed to be clearly in order for the function words.

The usage of three of the function words in the 100 word airline grammar also is worthy of discussion. The word "the" is always an optional one word node and may or may not be included in a given phrase. The words "of" and "for" constitute a two word node which appears in two separate paths of the syntax. These three words never affect the meaning of an airline query phrase. For example there is no semantic difference between the phrases "Report *the* current weather *of* San Diego" and "Report current weather *for* San Diego". We address this subject here because, in following chapters, we frequently present the word recognition accuracy for all words along with the word recognition accuracy excluding "the", "for", and "of".

6.3 Preliminary Experiments

For the first development experiment, template sets were made from the first vocabulary repetition for each non-function word and from an extracted template for each of the function words. For the second preliminary experiment, the three vocabulary repetitions were averaged to produce an "averaged" template set for each speaker. These templates were then employed in template extraction cf three tokens for each of the function words which were then averaged and substituted for their isolated counterparts in the averaged template sets. The recognition results for preliminary experiments are presented in Table 6-1.

Table 6-1
Results of Preliminary Experiments
For Ten Development Speakers.
494 Phrases from 100 word Airline Vocabulary

	All Words			Excluding "the,for,of"		
Template Set	Word Trials	Word Rate	Phrase Rate	Word Trials	Word Rate	Phrase Rate
Single tokens	3633	79.2	24.8	2854	85.7	53.8
3 Avg. Tokens	3633	84.9	39.8	2854	88.7	64.2
3 Avg. Tokens/Silence	3633	86.4	41.4	2854	90.6	67.0

Three types of word errors may occur in the recognition of a phrase: substitution of a wrong word for a spoken word, deletion of a spoken word, and insertion of a word which was not spoken. The first two errors cause a decrease in the count of correct words recognized. The insertion error is noted by increasing the count of total word trials. Thus, the word recognition rate is computed according to the following formula:

Word Rate = (100 x Correct-words) / (total-words + insertions)

In the first two experiments, it was noted that recognition errors were often introduced when the speaker paused in the midst of a long phrase. To correct this problem, a silence template was included in template storage for each speaker. The silence template is tested automatically for possible insertion between every word of the phrase and at the end of the phrase by the recognition algorithm. As noted in Table 6-1, the silence template improved overall word recognition accuracy by 1.5 percent. We decided to include the silence template in all subsequent experiments.

The experiment with three averaged templates reduced the word error rate by about one-fourth of the single token rate. Excluding "the", "for", and "of" errors, the word recognition rate was 90.6% versus 86.4 for all words. Thus, thirty percent of all word errors were the insertion or deletion of "the", or the confusion of "for" and "of".

6.4 Comparison of Training Techniques

The next series of development experiments was designed to evaluate the performance of extracted versus isolated templates in the recognition process. A set of 66 phrases had been recorded by each speaker for the purpose of template extraction. This set includes at least two occurrences of each vocabulary word. As described in Chapter 5 tokens were automatically extracted for each speaker with the recognition algorithm recognizing the phrases in a forced recognition mode using the three averaged tokens from the final preliminary experiment as templates.

Recognition accuracy was then compared over four template sets. The first template set was comprised of the three isolated averaged tokens used in the preliminary experiment. A second template set was created by averaging two extracted tokens for each word. The third experiment employed the union of the first and second template sets, i.e., two templates for each word. A fourth template set was made by averaging the two extracted tokens and the three vocabulary repetition tokens for each word. Results of these experiments are presented in Table 6-2.

Table 6-2
Results of Recognition Experiments
For Ten Development Speakers.
494 Phrases from 100 word Airline Vocabulary

Template Set	All Words		Excluding "the,for,of"	
	Word Rate	Phrase Rate	Word Rate	Phrase Rate
#1 Avg Three Isolated Tokens	86.4	41.4	90.6	67.0
#2 Avg Two Extracted Tokens	87.7	46.6	90.7	66.6
#3 Two Templates/Word: #1 and #2	91.8	58.9	95.6	81.8
#4 Avg Five Tokens	90.0	49.8	93.8	75.2

The results indicate little performance difference between averaged-isolated and averaged-extracted templates. Results with two templates per word significantly improved performance, cutting word recognition errors almost in half when the three function word errors are excluded. However, using two templates per word doubles both template storage and processing requirements and therefore is unfeasible in light of the real time response goal for the LCSE CSR system. Word accuracy with template set #4, the average of five tokens per word, was significantly better than either template sets #1 or #2 and, since set #4 uses only one template per word, it appears to be the most realistic training approach. A comparison of the word rate with template #4 and with the single token template set of the first preliminary experiment (Table 6-1) indicates that averaging five tokens cuts the overall word errors in half (accuracy changes from 79.2% to 90.0%). This figure is entirely consistent with the word error reduction on the digit phrase task in the template averaging study presented in Chapter 5.

In Table 6-3, we present the word error rate with template set #4, for six subsets of the vocabulary. The table clearly shows two sources of recognition errors, the function words and the digit-decade-teen group. The word rate on 1985 occurrences of the other 66 words in the vocabulary was 97.4%. Nearly all of the 66 word group are multi-syllable words including airline and city names and the 23 word alpha set. Averaging of tokens taken both from isolation and from continuous speech appears to be quite effective for the multi-syllable word group.

Table 6-3
Categories of Word Errors
With Averaged Templates From Five Tokens
For Ten Development Speakers.

CATEGORY*	TRIALS	ERRORS	WORD RATE
Function #1 ("the, for, of")	779	176	77.3%
Function #2 ("to, at, from")	260	39	85.0%
Digits ("0-9")	329	34	89.7%
Teens ("10-19")	164	16	90.2%
Decades ("20-90")	116	14	87.9%
66 Other words	1985	51	97.4%
TOTAL	3633	330*	90.9%

*Insertions not included.

6.5 Modification of Template Extraction Algorithm

In an effort to further improve performance, we examined the template extraction algorithm and found that the process was occasionally producing extracted tokens with faulty endpoints. To overcome this problem we added word score testing to the algorithm. If the normalized DPA score of the word preceding and following the word to be extracted were below a threshold, an extracted token was output. If the threshold test failed for either bounding word, extraction was not performed. The word score test was intended to insure that the speech frames to be extracted had been properly aligned by the dynamic programming algorithm.

A final development testing experiment was performed using the modified extraction algorithm. In this experiment, the number of extracted tokens per word was not limited to two While the standard set of 66 phrases contains each vocabulary word at least twice, commonly occurring words appear multiple times. In some cases, as many as nine tokens were extracted for a word. On the other hand, the word score test caused rejection of some extracted tokens and in some cases there were no extracted tokens available for the average template for a word. Averaged templates were generated by averaging the three vocabulary repetitions and all available extracted tokens. Results of the final development test experiment are presented in Table 6-4.

Table 6-4
Results of Final Development Test
For Ten Development Speakers.
494 Phrases from 100 Word Airline Vocabulary

	All Words		Excluding "the, for, of"	
Speaker Number	Word Rate	Phrase Rate	Word Rate	Phrase Rate
03	89.6%	49.0%	93.8%	69.4%
06	89.7%	49.0%	94.6%	77.6%
11	94.3%	61.2%	98.9%	93.9%
16	85.5%	32.7%	93.4%	73.5%
21	94.1%	67.3%	97.4%	89.8%
23	94.5%	65.3%	96.9%	83.7%
31	96.6%	76.0%	99.6%	98.0%
36	96.3%	76.0%	99.0%	96.0%
41	92.1%	62.0%	95.7%	82.0%
43	92.2%	54.0%	97.3%	88.0%
Average →	92.5	59.3	96.7	85.2

Comparison of the recognition rates in Table 6-4 with those in Table 6-2 for template set #4 shows that significant improvement was obtained by the modified template extraction algorithm. Overall word rate improved from 90.0% to 92.5%, while the phrase rate improved from 49.8% to 59.3%. Word rates for individual speakers ranged from 85.5% to 96.6%.

6.6 Training Technique for Performance Tests

After evaluation of the various training approaches for the development speakers, we established the training procedure for the 50 speaker performance test. The final approach is a five-step automatic process as follows:

1. Average the three vocabulary repetitions for each word in the 100 word vocabulary.
2. Using the templates created in Step 1, extract multiple tokens for the six function words ("of, for, the, to, at, from") from the standard phrase set for each speaker.
3. Average the extracted tokens for each of the function words.
4. Using the function word templates from Step 3 and the remaining templates from Step 1, extract multiple tokens for all words from the standard phrase set.
5. For each word, average all available tokens, thus generating the template to be used in the performance test. For non-function words, the available templates include the three vocabulary repetitions and all tokens extracted in Step 4. For function words, the vocabulary repetitions are excluded from input to the final average template.

7. PERFORMANCE TEST

The performance tests were carried out on a data base of 50 random phrases from the airline query grammar spoken by the 50 data base subjects. Each of the subjects also spoke a training set consisting of three repetitions of the 100 word airline vocabulary and 66 phrases used for template extraction. A single averaged template for each vocabulary word was generated for each speaker according to the procedure described in Section 6.6.

7.1 Performance Test Results

A summary of performance test results is presented in Table 7-1. The average word recognition rate for all words was 93.1%, while excluding "of", "for", and "the" errors, the average word rate was 95.7%. The average phrase recognition rate was 64.6% and the corresponding rate was 83.0%, when "of", "for", and "the" errors are ignored. A phrase is considered correct if all words in the phrase are correctly identified. Note that of the 884 phrases which were recognized incorrectly, in 459 cases, the only error was the confusion of "of" and "for" or the deletion or insertion of "the".

On each recognition trial, the CSR system reports five candidate recognition results ranked by score. In Table 7-1 the rows labelled "OPTION" show the number of times each candidate was the correct result. The line labelled "OPTION #2" indicates that in 310 cases (12.4%), the CSR system's second choice was the correct phrase.

Table 7-1
Summary of Performance Test Results
100 Word Airline Grammar

50 Speakers - 50 Phrases per Speaker

ALL WORDS			EXCLUDING "of, for, the"	
PHRASE TRIALS =	2500			
CORRECT =	1616	64.6%	2075	83.0%
OPTION #2 =	310	77.0%	149	89.0%
OPTION #3 =	75	80.0%	35	90.4%
OPTION #4 =	38	81.5%	12	90.9%
OPTION #5 =	24	82.5%	12	91.3%
MEDIAN PHRASE RATE	68.0%		84.0%	
WORD TRIALS =	18416		14873	
CORRECT =	17284		14307	
INSERTIONS =	144		77	
DELETIONS =	358		61	
WORD RATE =	93.1%		95.7%	
MEDIAN WORD RATE =	94.5%		96.8%	

A complete tabulation of phrase and word recognition results by individual speaker is included in Appendix C. Figure 7-1 presents a histogram of word rates and phrase rates for all 50 speakers. The median of the distribution of overall word rates is 94.5%. The median figure is 1.3% higher than the overall average word rate for all speakers. The median provides a better estimate of the expected performance of an unknown speaker, because as Figure 7-1 shows, the average word rate is lowered significantly by a small group of poor performing speakers. Excluding "of", "for", and "the" errors, the median rate is 96.7%, one percent higher than the corresponding average word rate.

In Table 7-2, we present a summary of word recognition rates according to the sex, age, and educational background of the speakers. The overall average rate of female speakers exceeded that of males by .8%. The age summary suggests older speakers perform better than younger while educational background seems to have no correlation with word rate performance.

FIGURE 7-1

PERFORMANCE HISTOGRAMS
50 SPEAKERS
EXCLUDING OF, FOR, AND THE ERRORS

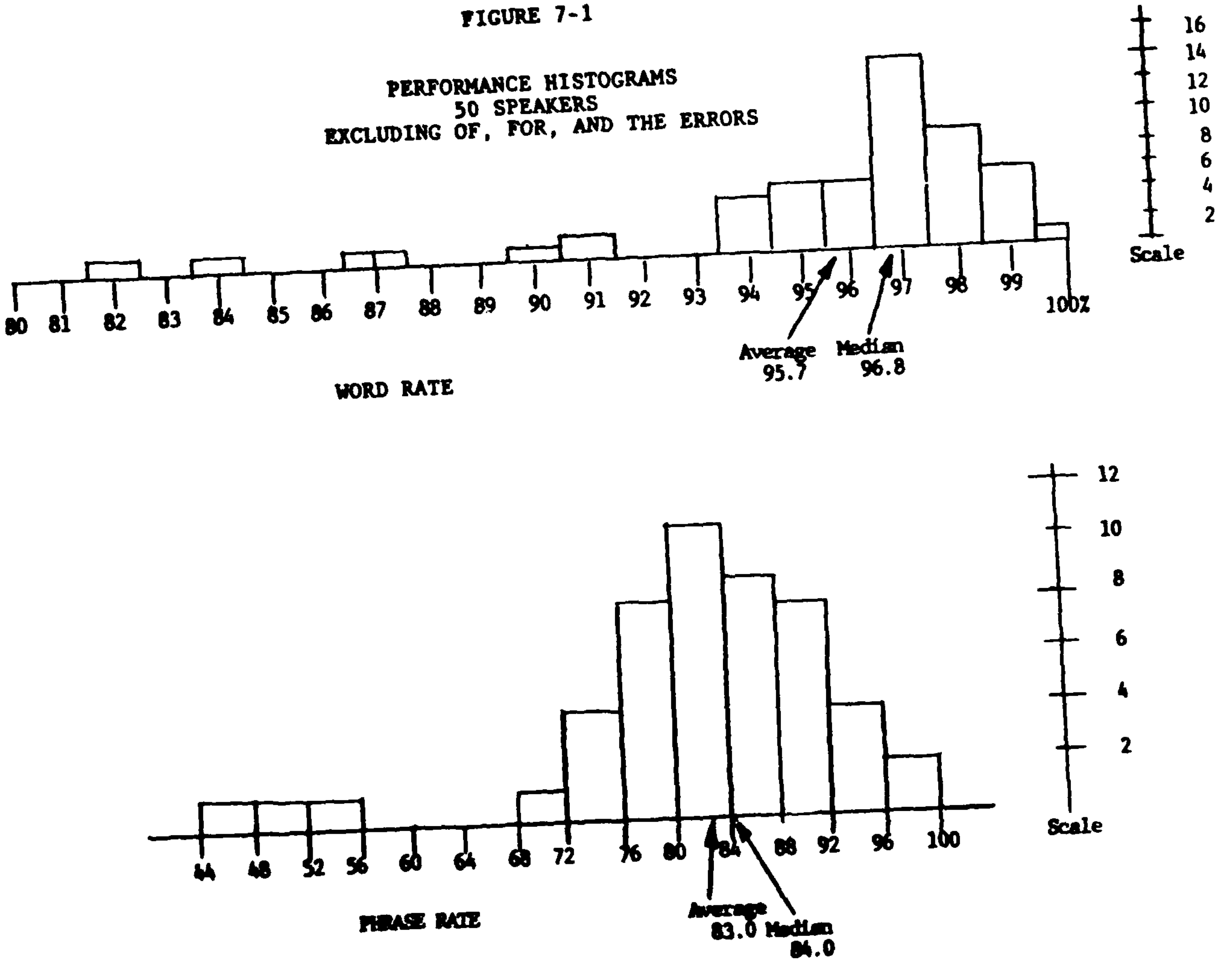


Table 7-2
Summary of Word Recognition Rates by Speaker's
Sex, Age and Educational Background.

SEX

Male	Female
92.7%	93.5%

AGE

	Teens-Twenties	Thirties	Forties-Fifties
Number of Speakers	28	15	7
Overall Word Rate	92.5%	93.4%	94.8%

EDUCATION

	High	Junior	Bachelor	MS	
	School	College	Degree	Degree	Ph.D
Number of Speakers	8	16	16	7	3
Overall Word Rate	93.3%	92.7%	94.0%	91.7%	93.4%

7.2 Categories of Performance Test Errors

Table 7-3 presents word recognition rates for six subgroups of the 100 word vocabulary. These figures show that the CSR system has particular difficulty identifying the function words, the digits and the decades. The function words are frequently deemphasized in continuous speech while the digits and decades are frequently confused with each other. The word rate for the 66 word group, which makes up the majority of the word trials is 98.4%. 63 of the 66 words in this group are multi-syllabic.

Table 7-3
Word Recognition Rate by Vocabulary Subgroups

Category	Trials	Word Rate
"of, for, the"	3543	82.4%
"to, at, from"	1738	92.5%
"digits"	2458	89.9%
"teens"	441	94.2%
"decades"	416	82.8%
"66 remaining words"	**9820**	**98.4%**
All Words	18416	93.1%

7.3 Performance Test With 82 Word Vocabulary

In order to gain insight as to the performance of the CSR system on a less challenging syntax, another performance test was designed and conducted with an reduced vocabulary. The decade and teen nodes were eliminated from the syntax, thus reducing the vocabulary to 82 words. The templates for each speaker were the same as those used in the 100 word test. The test phrases were a subset of those used in the 100 word experiments and were obtained by eliminating all airline phrases containing a teen or decade word. This reduced the average number of test utterances per speaker from 50 to 32. Table 7-4 shows a comparison of performance on this subset of the airline phrases. Results are given first with the 100 word vocabulary, including the decades and teens and then with the reduced 82 word vocabulary.

Table 7-4
Comparison of Performance of
100 Versus 82 Word Airline Vocabulary

50 speakers
1650 Phrase Trials

	All Words		Excluding "of, for, the"		
Vocabulary	Median Phrase Rate	Median Word Rate	Median Phrase Rate	Median Word Rate	Digit Word Rate
100 Words	71.0%	94.6%	84.9%	96.9%	89.8%
82 Words	74.2%	95.8%	90.5%	**98.0%**	94.2%

When the vocabulary was reduced from 100 to 82 words, the median word rate improved from 94.6% to 95.8%. Excluding "of", for", and "the" errors, note that a 98.0% word rate (or 2.0% error rate) was achieved on the 82 word test while the word error rate was 3.1% for the 100 word vocabulary. Elimination of the decades and teen reduced word errors by 35.0%.

The corresponding figures for phrase rate errors (excluding "of", "for", and "the") were 9.5% and 15.1%. Thus, the number of phrase errors declined 37% with the reduced vocabulary. The 82 word experiment demonstrates the impact that syntax and vocabulary selection can have on the performance of a CSR system.

The right hand column of Table 7-4 shows the word recognition rates for the ten digits. In the 82 word experiment, the digits could not be confused with teen and decade words and 43% of the digit word errors were eliminated.

8. ERROR ANALYSIS

Error analysis in the context of this report is an attempt to classify the causes of the recognition errors made by the CSR system in performance tests. The data base for the error analysis is made up of those performance test phrases which were incorrectly identified. Errors on the semantically irrelevant function words "for", "of", and "the" were ignored in this study. There were 425 of the 2500 performance test phrases (i.e., about 17%) that contained word errors other than "for", "of", and "the". These 425 airline query phrases constitute the data base for error analysis in this report.

8.1 Preliminary Analysis

The first step was to gather statistics on word and phrase trials, insertions, deletions, and substitutions for individual speakers and for the group of 50 speakers. Word accuracies were compiled for various subsets of the 100 word vocabulary such as function, words, digits, teens, and decades. Many of these statistics were presented in the tables of Chapter 7. A more detailed account of individual word errors in the vocabulary subgroups is presented in Table 8-1. The function words aside, the most commonly misrecognized words were the digit "eight" which varies according to the presence of the stop release and the decade "seventy" which is often mistaken for the digit "seven".

A list of all phrases in error was compiled and for these phrases, the scores and endings of individual words were obtained. Data was also gathered for each speaker's templates. This data included duration of the averaged template, the number of isolated and extracted tokens which made up the averaged cluster, and whether the cluster center token was isolated or extracted.

Table 8-I
Recognition Statistics
By Individual Word

	Occ	Ins	Del	Sub	Sub*	Rate
the	1521	70	296	14	0	79.6%
for	965	0	0	121	139	87.5%
of	1057	0	0	135	127	87.2%
TOTALS	3543	70	296	270	266	82.4%

	Occ	Ins	Del	Sub	Sub*	Rate
from	806	5	1	35	61	95.5%
to	887	2	2	70	24	91.9%
at	45	3	1	12	10	71.1%
TOTALS	1738	10	4	117	95	92.5%

	Occ	Ins	Del	Sub	Sub*	Rate
zero	95	2	0	4	2	95.8%
one	276	1	3	10	18	95.3%
two	321	10	6	13	19	94.1%
three	276	1	4	17	23	92.4%
four	256	2	0	25	12	90.2%
five	212	3	1	18	8	91.0%
six	273	1	2	19	4	92.3%
seven	223	2	0	25	16	88.8%
eight	245	12	15	42	8	76.7%
nine	281	0	2	11	8	95.4%
TOTALS	2458	34	33	184	117	89.9%

	Occ	Ins	Del	Sub	Sub*	Rate
ten	15	3	0	0	5	100.0%
eleven	46	1	0	1	0	97.8%
twelve	35	0	0	0	3	100.0%
thirteen	29	0	0	1	0	96.6%
fourteen	44	0	0	1	4	97.7%
fifteen	54	1	0	3	4	94.4%
sixteen	59	0	0	4	7	93.2%
seventeen	61	0	0	7	3	88.5%
eighteen	48	0	0	4	3	91.7%
nineteen	50	0	0	0	3	100.0%
TOTALS	441	5	0	21	32	94.2%

	Occ	Inc	Del	Sub	Sub*	Rate
twenty	61	0	0	12	5	80.3%
thirty	73	0	0	13	5	82.2%
forty	70	0	0	9	25	87.1%
fifty	41	0	0	7	10	82.9%
sixty	52	0	1	5	20	88.5%
seventy	60	0	0	14	26	76.7%
eighty	28	1	0	6	25	78.6%
ninety	31	1	0	3	5	90.3%
TOTALS	416	2	1	69	121	82.8%

	Occ	Ins	Del	Sub	Sub*	Rate
report	369	0	2	7	2	97.6%
what-is	368	0	0	8	8	97.8%
tell-me	361	0	1	1	2	99.4%
give-me	383	0	0	1	7	99.7%
arrival-time	218	0	0	0	1	100.0%
departure-time	266	0	0	2	0	99.2%
location	248	0	0	0	1	100.0%
status	259	0	0	0	2	100.0%
flight-schedul	406	0	0	1	3	99.8%
aircraft-type	265	0	0	2	0	99.2%
passenger-load	271	0	0	0	0	100.0%
distance	160	0	0	0	0	100.0%
flight-time	169	0	0	2	1	98.8%
current	103	1	0	0	2	100.0%
weather	190	0	0	1	0	99.5%
forecast	143	0	0	0	0	100.0%
flight	1160	0	1	18	13	98.4%
number	290	1	0	7	1	97.6%
aircraft	260	0	0	0	7	100.0%
hundred	86	2	2	0	1	97.7%
national	188	0	0	1	1	99.5%
twa	129	0	0	0	0	100.0%
united	134	0	1	0	1	99.3%
psa	146	0	0	0	1	100.0%
western	119	0	0	1	2	99.2%
continental	159	0	1	6	1	95.6%
american	157	0	0	1	2	99.4%
hughes-airwest	124	0	0	5	0	96.0%
eastern	130	0	2	1	2	97.7%
allegheny	167	0	0	1	1	99.4%
los-angeles	226	0	0	3	0	98.7%
san-diego	160	2	0	2	0	98.8%
washington	199	0	0	2	0	99.0%
denver	221	0	3	4	6	96.8%
dallas	175	0	1	4	2	97.1%
new-york	176	1	3	3	2	96.6%
chicago	231	1	2	4	2	97.4%
boston	212	0	1	4	1	97.6%
atlanta	204	0	4	7	1	94.6%
pittsburgh	173	1	0	1	2	99.4%
alpha	32	0	0	6	1	81.3%
bravo	10	0	0	0	2	100.0%
charlie	19	1	0	0	2	100.0%
delta	31	1	0	2	4	93.5%
echo	5	1	0	0	3	100.0%
foxtrot	14	0	0	0	1	100.0%
golf	6	1	0	0	3	100.0%
hotel	7	0	0	0	0	100.0%
india	16	0	0	3	0	81.3%
juliet	11	1	0	0	4	100.0%
kilo	45	1	0	1	2	97.8%
lima	4	0	0	1	9	75.0%
mike	11	3	0	0	12	100.0%
november	24	1	0	0	2	100.0%
oscar	16	1	0	0	1	100.0%
papa	26	0	0	0	7	100.0%
quebec	6	0	0	0	1	100.0%
romeo	21	0	0	0	0	100.0%
sierra	11	1	0	0	1	100.0%
tango	27	0	0	0	3	100.0%
uniform	5	0	0	0	4	100.0%
victor	27	0	0	0	8	100.0%
whiskey	7	1	0	0	0	100.0%
x-ray	10	0	0	0	2	100.0%
yankee	20	1	0	0	3	100.0%
zulu	27	0	0	0	3	100.0%
TOTALS	9820	23	24	113	146	98.4%

* word was substituted for another word

8.2 Listening Procedures

Two modes of listening were employed in error analysis. In the first mode, an audio signal was synthesized from the filter coefficients contained in test utterance files. This mode was valuable for verifying the end point detection process. The second listening mode was to playback the recorded audio tape for individual speakers for purposes of phonetic comparison. Listening procedures were used only for those phrases containing multiple word errors or suspected end point detection problems.

8.3 Error Classification and Coding

Table 8-2 contains a list of ten classes of CSR recognition errors including a class for errors whose cause is unknown and a class for errors whose cause requires further analysis. The ten classes often contain subgroupings which identify the cause of the error more specifically.

The phonetic similarity class applies to those words which were mistakenly identified as a similar sounding word(s). The end point detection class contains those errors in which the CSR system did not properly detect the beginning or end of the phrase. Template generation classifies those errors which were clearly due to an inadequate template for the word. Pronunciation errors are a self-explanatory class.

Error classification #4 is labelled "pause in unknown for multisyllable word". The CSR algorithm is adept at handling pauses between words, but may make an error when a pause occurs within a word such as in "con-(pause)-tinental".

The "gap between words" category applies to pauses between words which are too short to be recognized by the silence template. This type of error occurs most frequently in the "digit-decade-teen" portion of the airline syntax. The "1 to N" category refers to recognition errors such as "seventeen" being identified as "seven

Table 8-2
TABLE OF CODES FOR CAUSES OF CSR-SYSTEM ERRORS

0 CAUSE UNKNOWN 0 = 53

1 PHONETIC SIMILARITY 1 = 315
- a. Whole-word similarity (five/nine)
- b. Partial-word similarity (sixteen/six)
- c. Word boundary crossing (seventeen/seven ten)
- d. Coarticulation (two eight/three)

2 ENDPOINT DETECTION 2 = 8
- a. Beginning of sentence
- b. End of sentence

3 TEMPLATE GENERATION 3 = 145
- a. Isolated template too long--no extracted tokens
- b. No extracted tokens for some other reason
- c. Cluster center has extreme duration
- d. Variation in final-stop release
- e. Difference in stress patterns between isolated and extracted tokens
- f. Difference in intonation between isolated and extracted tokens
- g. Template is too short
- h. Template is too long

4 PAUSE IN UNKNOWN FOR MULTISYLLABLE WORD 4 = 9

5 GAP BETWEEN WORDS 5 = 56
- a. digits, teens, or decades
- b. Other words

6 PRONUNCIATION ERROR 6 = 57
- a. Dropping amplitude at end of sentence
- b. Pause
- c. Stuttering
- d. Excessive reduced duration
- e. Excessive extended duration

7 1-TO-N OR N-TO-1 ERROR 7 = 86
- a. n to 1
- b. 1 to 2
- c. 1 to 3
- etc.

8 PROPAGATION ERRORS 8 = 153
- a. Adjacent-word error and syntax constraint
- b. Nonadjacent-word error and syntax constraint
- c. Adjacent word has bad word boundary
- d. Illegal syntax

9 FURTHER ANALYSIS REQUIRED 9 = 111

Total = 639

ten" while the "N to 1" category is the reverse.

The classification "propagation error" contains those word errors caused by the misrecognition of a previous word in the sentence. This type of error results from the use of syntax to restrict the potential word candidates in various parts of the sentence. Occasionally, if a key syntactic word is misidentified in the phrase, the proper templates are not matched with the speech following this word.

To illustrate, consider the phrase "Tell-me the status of aircraft alpha bravo thirty two". In the airline query grammar, the words in the "alpha" node must be preceded by the word "aircraft". If "aircraft" is not identified, the words "alpha" and "bravo" will also be misidentified. Errors such as "alpha" and "bravo" in the above example would be classified as propagation errors while "aircraft" would be placed in another error category.

8.4 Results of Error Analysis

Certain word errors cannot be ascribed to a single cause. For example, a word may be identified as a similar sounding word, but may also have an inadequate template or may cause a one-to-N or N-to-one type error. Thus, there is not a one-to-one correspondence between word errors and error classifications.

The data base of 425 phrases with errors contained 639 individual word errors. A summary of the classification of these errors is presented in Table 8-2. The phonetic similarity class contains 315 errors. Thus, for nearly half of the words in error, the CSR system recognizes a similar sounding word or words.

The second most common error category is propagation errors. In these cases, correction of the source of the propagation error would likely correct multiple errors. The third most common category is template generation. Improved training techniques may correct many of these word errors and errors in other classifications as well.

The most challenging aspect of the 100 word airline query grammar is recognizing the phonetically similar digit, teen, and decade words. A confusion matrix of errors between these words is presented in Table 8-3. Propagation errors are not included in this tabulation. The rows in Table 8-3 correspond to the intended word while the columns represent the mistakenly recognized word. At the far right of the table is a column containing the number of word trials, the number of times each word appeared in a performance test phrase. The diagonal boxes in the table denote the confusions of each word with its phonetically similar counterpart, for example, "four" with "forty" or "thirty" with "three". The confusion of "seven" as "seventy" occurred 23 times in the performance test while "eight" was confused with "eighty" on 17 occasions. The clustering of the confusions on the diagonals clearly demonstrates the difficulty of recognizing competing similarly sounding words.

Table 8-4 presents a tabulation of the number of words in error for each of the 2500 phrases of the performance test. 2085 phrases had no errors while 294 had just one error. The second line of this table presents the data where propagation errors are ignored. Propagation errors aside, in 85% of the cases where the CSR system misrecognized a phrase, it misrecognized only one word in the phrase. This indicates that CSR system word errors are somewhat independent with the exception of propagation errors, of course.

TABLE 8

Confusion Matrix of Digits, Teens, Decades (without propagation errors)

	Recognized Word																															Word
Intended Word	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	30	40	50	60	70	80	90	Other	Ins.	Del.	Trials
0																													4			95
1					3																		2						2		2	276
2							1		1	1	1						1				1			1	2				1	8	3	321
3			6					1														2	2	1					2		2	276
4		6							1				1		1						1		10						2	1		256
5		1			1					3			1								1							1	6		1	212
6			1	1					1										1						12				2	2	2	273
7																					1					23			1		1	223
8			4	6															2		1	1		1			17		3	9	13	245
9		4				1																						2	1			281
10																														3		15
11																															1	46
12																																35
13																1																29
14					2																											44
15																								3						1		54
16																									4				1			59
17								4																		3						61
18																1				1				1								48
19																												1				50
20		3			1																		7									61
30				12																				1								73
40				1	6										2																	70
50																2	1								1				2	1		41
60							1										5															52
70								13									1	3														60
80									3										3				1							1		28
90										1										1												31

Table 8-4

Number Of Word Errors Per Phrase

	0	1	2	3	4	5 or more
WITH PROP. ERRORS	2085	294	73	22	11	15
WITHOUT PROP. ERRORS	2085	346	57	5	1	0

9. CONCLUSIONS

In performance of the Limited Connected Speech Experiment, ITTDCD has demonstrated that a one hundred word vocabulary can be processed in real time with high word recognition accuracy while concurrently providing a flexible "Voice-control" feature which allows the user to control the CSR system via spoken commands of his own choosing.

Three major conclusions from this contract are presented below:

1. The tem‚ ate averaging study discussed in Chapter 5 showed that a speaker dependent CSR system with storage and processing limitations achieves better performance with a set of averaged tokens than with any individual set of tokens.

2. Performance of the CSR algorithm is extremely dependent on the quality of word templates. In the initial development experiment discussed in Chapter 6, a word rate of 85.7% and a phrase rate of 53.6% was obtained with isolated single word templates. For the final development experiment the corresponding figures for the same test phrases and speakers were 96.7% and 85.2%. The recognition algorithm was the same for both experiments but the training technique was modified.

3. CSR performance on a given task will vary sharply according to the structure of the task synta˜ and the choice of the task vocabulary In the 100 word versus 82 word experiment described in section 7.3, a

median word rate of 94.6% was achieved for all words on the 100 word test. For the same test phrases and speakers on the 82 word test, the median word rate was 98.0% (excluding "of,for,the" errors). A reduction in word errors of 63% was thereby achieved by restructuring the syntax, removing phonetically similar words from the vocabulary, and ignoring semantically irrelevant errors.

APPENDIX A

CONNECTED SPEECH RECOGNITION SYSTEM

USER'S GUIDE

User Interface

That portion of the VAX-11/780 CSR system which is visible to the user is called the user interface. Through this interface, the user invokes the CSR program and indicates to it the type and order of operations to be performed.

1. Program Invocation

The VAX-11/780 CSR system is much like any other UNIX© applications program. It can be executed any time when the user is at the shell level, although it is not currently linked to either /bin or /usr/bin. This requires that an absolute pathname be used to start up the program. That pathname currently is

/usr1/a/sr/bin/csr

The CSR system accepts optional switches at invocation which affect its processing. The syntax of the program call is

csr [-ceistv] [*extension* [*command_file* ...]]

(This assumes that a *chdir*(1) to the resident directory has been done previously.) The switches include

-c Continue processing following fatal errors. This is the default setting. Be careful, however, since fatal errors can cause spurious results to occur. This switch is mainly intended for program testing and short program runs without command files. Long experiments should use the -t switch (see below).

-e Use the next program argument as an extension when forming both the DPA analysis and experiment results pathnames. In both cases, the extension (or as much of it as will fit) is appended onto the end of the formed pathname. Keep in mind that UNIX pathnames are limited to 14 characters.

-i Open a file with the name *.csrrc* in the current directory and perform its commands as if they had been entered from the keyboard. This is useful for performing repetitive initialization sequences. Normal processing continues following end-of-file on the initialization file.

-s Operate in silent, rather than verbose, mode. Normally, each CSR command displays terse information concerning its execution. This information can be suppressed by specifying this switch. Fatal errors and program diagnostics cannot be suppressed, only informative messages.

-t Terminate on first occurence of a fatal error. This should always be used for long experiment runs, since an error can result in spurious results from that point on.

-v Operate in verbose, rather than silent, mode. This is the default setting. Normally, each CSR command displays terse inforamtion concerning its execution. This information is written to the user's terminal or to the experiment results file (if experiment mode is on). Fatal errors and program diagnostics are always written to the standard output unit and cannot be suppressed or redirected without use of the shell's redirection facilities.

If command file(s) are specified on the invocation line, the program automatically terminates when the end of the last command file is reached, provided no fatal errors or other terminating conditions were encountered. Otherwise, the program will prompt the user for valid commands and terminate upon entry of the quit command.

2. Program States

While the VAX-11/780 CSR system is running, it is in one of three distinct states or modes:

- **Command** mode is the normal, default state of the CSR system. In this mode, commands are read from either the keyboard or specified command files. Each recognition trial must be explicitly performed by issuing the *recognize_speech* command. Whenever an interrupt is caught, the system returns to this state, regardless of what state it was in before the interrupt.
- **Ongoing Recognition** mode is the program state in which recognition trials occur one after another with no intervening user action. As soon as the results of one recognition trial are displayed, the system is *listening* for the next unknown utterance. This mode is entered from the Command mode by setting the *Single_recog* switch off.

  The user may exit to the Command mode by hitting an interrupt (labeled **DEL** or **RUBOUT** on some terminals) or to the Voice Control mode by saying the isolated word *control*. While in Ongoing Recognition mode, there is no input read from the terminal or any open command file(s).
- **Voice Control** mode is made the current state if the user says the isolated word *control* or types it in while in Command mode or if the isolated word *control* is spoken while in Ongoing Recognition mode. A small subset of the commands that the user has available in Command mode are available in Voice Control mode. The main difference is that while in Voice Control mode all commands are spoken, rather than typed in from the keyboard. Typical commands are isolated words that require no arguments.

  The user may exit to the Command mode by hitting an interrupt or by saying *offline*. Similarly, saying *recognize* switches the system into Ongoing Recognition mode. While in Voice Control mode, there is no input read from the terminal or any open command file(s).

  NOTICE: Although being in Voice Control mode requires ongoing recognition, it is quite different from the Ongoing Recognition mode in that only the control syntax/vocabulary is active. Ongoing Recognition mode, on the other hand, generally has the non-control syntax/vocabulary active (with the exception of the meta-control words *cancel* and *control*).

3. Commands

3.1. Command List

The list of valid CSR commands includes commands which are allowed only from the keyboard (preceeded by the † symbol), allowed only from a command file (preceeded by the ‡ symbol) and those allowed both from the terminal and spoken in Voice-control mode (preceeded by the • symbol). Valid commands from the keyboard are entered in response to the system's prompt (which is initially ">", but can be changed under user control). All command names can be abbreviated to as few characters as are necessary to guarantee uniqueness. For example, in the list below *sp*, *speak* and *speaker* are all valid names for the same command. However, specifying a command name of *s* is ambiguous because *save_speech*, *set_environment*, *signal* and *summary*, as well as *speaker* all begin with the letter *s*.

analysis [off | on [*analysis_results_pathname*]]
average [*averager_arguments* ...]
banner [*one_line_message*]
change_directory [*new_working_directory*]
clear_template *word* [*directory*]
command_file *command_file_pathname*
• **control**
• **display**
† **document** [*command_name* ...]
experiment [off | on [*experiment_results_pathname*]]
extract [fct] [*extractor_arguments*]
† **help** [*command_name* ..]
live_experiment *standard_phrase_pathname*
load_syntax *syntax_specification_pathname*
load_templates
• **offline**
• **options**
phrase_training [*starting_phrase_number*] [*output_directory*]
pwd
quit/^D (control-D)
• **recognize** [*digital_unknown_pathname*]
reset_environment [*variable_name* ...]
save_speech *output_pathname*
set_environment [*variable_name new_value*] ...
signal *UNIX_signal_name* [off | on]
speaker *initials*
summary [*one_line_message*]
task *task_name*
train_system [*utterance_string* [*output_directory*]]
‡ **tty_input**
unix [*C_shell_command*]
version
vocabulary_training [*beginning_word*] [*repetition_count*]
• **word-scores**
! [*C_shell_command*]
[*one_line_comment*]

3.2. Command Descriptions

3.2.1. **analysis** [off | on [*analysis_results_pathname*]

This command changes the current state of analysis processing. When turned on, detailed recognition information that includes DPA distances, directions of movement through the DPA matrix and frame-by-frame recognition scores. In either the off or on setting, analysis processing does not affect the recognition algorithm.

When used with no arguments, the current state is toggled (from off to on, or vice-versa). If only one argument is present, it must be either the string *off* or *on*. When turning analysis mode on in this case, the previously-used or default filename is opened as the output file. The default pathname is of the form

/tmp/*llll.pppppp*[*ee*]

where "*llll*" is the first four letter of the user's logon name, "*pppppp*" is the zero-filled, six-digit process ID number (of the object program), and "*ee*" is an optional extension that the user specified at program invocation by using the **-e** switch. If the user specifies a pathname when turning the analysis mode on, that pathname takes precedence over the default name.

3.2.2. **average** [*averager_arguments*]

The average command executes the template averager and, with no passed arguments, produces averaged templates in the speaker's *a02* directory. The averager only uses those files that have been created by training since the last averager run. Any arguments following the command name are passed in to the template averager. Since this command executes another UNIX program via *fork/exec*, it could take a few minutes to finish.

3.2.3. **banner** [*one_line_message*]

The banner command allows the user to conspicuously display a single line of text on the standard output unit. The message is preceeded and followed by three pound signs (#) and the message is optional. This command is useful for informing the user of events, such as the completion of recognition trials for a speaker.

3.2.4. **change_directory** [*new_working_directory*]

This command allows the user to change the program's concept of its current working directory. This is useful since speech files reside in many different directories. The full power of the shell's ***chdir***(1) command is supported. If no argument is specified, the directory which the user was in when the CSR system was invoked is returned to. If an argument is present, it is a pathname of a directory to change to. The specified pathname, and those higher in the hierarchy, must be searchable by the user and must be a directory. This command can also be abbreviated to **cd** or **chdir**. One note of caution: just as in the shell version of the command, a directory changed to is the current working directory until another instance of the command. Since many CSR commands take pathnames as arguments, be extremely careful when using non-rooted pathnames, since they will be affected by the current working directory and its location in the file hierarchy.

3.2.5. **clear_template** *word* [*directory*]

The clear_template command allows the user to discard all or some of the repetitions of a trained vocabulary word. The first argument is always required and is the vocabulary word whose template(s) need to be retrained. The second argument is an optional directory from which to clear the word. If the directory is not specified, the word is cleared from all directories owned by the user.

Recall that when in training, template files are created read-only. This effectively prohibits accidental destruction through retraining. If a retraining of a word is desired, the word's template(s) must be cleared first and then retrained. Since ***clear_template*** accomplishes its task by merely changing the access mode of the template file(s) to read/write by all, a cleared template is not destroyed until a retraining is done.

3.2.6. **command_file** *command_file_pathname*

This command is extremely useful for long or repetitive sequences of CSR commands. Commands are entered into a file with the assistance of one of the many available text editors. Then, each time this command is executed the commands in the specified file are performed as if they were entered by the user from the keyboard (with the exception of those keyboard-only commands). Command files may be nested to a depth of 9, which should be more than adequate for most applications. This number can be raised to a maximum of 15, if absolutely necessary by changing the value of the defined constant CFM_NEST to 15.

3.2.7. control

The *control* command places the CSR system into Voice Contro! mo 'e (see §2) Once in Voice Control mode, command input from the terminal and any open command files is suspended. There is a small subset of commands that are operative during Voice Control mode. These include:

- display
- offline
- options
- recognize
- word-scores

This command can being entered from the terminal keyboard or spoken during a recognition trial. The user can exit from Voice Control mode by saying *offline* or by hitting an interrupt. In either case, the program returns to Command mode.

3.2.8. display

This command displays the current values of the CSR system's settable engineering parameters. It is equivalent to typing *set* with no arguments. It is valid both as a keyboard entry and as a spoken command while in Voice Control mode.

3.2.9. document [*command_name* ...]

This command, when entered from the keyboard only, allows the user to peruse various external documentation files concerning valid CSR commands. If no command names are specified, the entire list of command documentation is shown to the user, in alphabetical order, one page at a time. If command name(s) are specified, only those commands are documented. In either case, the program fork/executes a UNIX shell to page through the output using the *more*(1) command. The external documentation is located in /usr1/a/sr/csr/csr.doc.

3.2.10. experiment [off | on [*experiment_results_pathname*]]

The experiment command changes the current state of experiment processing. When turned on, recognition results, as well as verbose program output, are written to the experiment results file instead of to the standard output unit.

When used with no arguments, the current state is toggled (from off to on, or vice-versa). If only one argument is present, it must be either the string *offR or on*. When turning experiment processi.g on in this case, the previously used or default filename is opened as the output file. The default pathname is of the form

/tmp/E*llll.pppppp*[*ee*]

where "*llll*" is the first four letters of the user's logon name, "*pppppp*" is the zero-filled, six-digit process ID number (of the object program), and "*ee*" is an optional extension that the user specified at program invocation by using the **-e** switch. If the user specifies a pathname when turning the experiment processing on, that pathname takes precedence over the default name.

3.2.11. help [*command_name* ...]

This command, when entered from the keyboard only, allows the user to view one-line summaries of valid CSR commands. If no command names are specified, the entire list of command summaries is shown to the user, in alphabetical order, one page at a time. If command name(s) are specified, only those commands are summarized. For more detailed documentation, use the **document** command.

3.2.12. **linear_transform** [off | on [*linear_transformation_pathname*]]

The linear_transform command changes the curent state of linear transformation processing. When turned on, all subsequent speech input is transformed by the linear transformation matrix that was last specified.

When used with no arguments, the current state is toggled (from off to on, or vice-versa). If only one argument is present, it must be either the string *off* or *on*. When turning linear transformation processing on in this case, the previously used filename is used as the transformation. The default transformation pathname is undefined (null) and thus the user must either use the **set_environment** command to set the *Transform* environment value to the appropriate transformation pathname or the linear transformation pathname must be explicitly specified the first time.

See §1.2 in Data Descriptions for more details.

3.2.13. **load_syntax** *syntax_specification_pathname*

The load_syntax command sets up the syntax data structures that are necessary to drive the recognition algorithm. The DPA process is syntax-directed and will not work without some kind of syntax present. Currently, for simple grammars that operate without syntax (for example continuous digits) there is a dummy syntax available in /speech/csr/syx/anyd.syx that permits any combination and order of vocabulary words. More complex grammars, not surprisingly, have more complex syntax specifications.

3.2.14. **quit**/^D (control-D)

The quit command is the only way to gracefully exit the program explicitly. If experiment processing is on, it is turned off and results are summarized prior to program exit. An alternate and equivalent way to quit the program is to enter a control-D.

3.2.15. **read_template** *speech_file_pathname* [*template_#*] [nosyntax]

This command allows the user to load template memory and to indicate the template number to be assigned and whether the template is constrained by the syntax. It and **downline_load** are the only ways to load template memory.

When only a single argument is present, it is the name of a UNIX file that contains speech written as described in §1.1 of Data Descriptions. The source of the data can be either PDP-11/60 or VAX-11/780 which is indicated by the *Source* environment value. The file is opened, the data is read into buffers and the buffers are pre-processed before being stored into template memory. This pre-processing is necessary due to the differences in how network information is stored. In the single argument case, the first available template number (starting at 1) is assigned. If a second argument is present, it is assumed to be the template number or the string "nosyntax" (or "nosyn" for short). In the former case, the specified number is assigned to the template and any existing template with that number is overwritten. If the "nosyn" string is present as the final argument, it indicates that this template is not constrained by syntax; i.e. the template can match any unknown utterance at any point in the partial phrase. A typical use for the *meta-syntax* indication is for the silence template. It is not part of the grammar, vocabulary or syntax, but it should be considered a candidate for matching within the DPA process at any time.

Although template numbers can be specified when reading speech templates into memory, it is advised against doing so. This is due to the restrictions that the syntax imposes. Templates must be loaded in the same order that the syntax specification is laid out. The DPA process expects template numbers to begin with 1 and increment with no "holes", or missing templates. If a template number is expected to be filled with a template and it is empty, the DPA process will fail to work properly.

3.2.16. **recognize_speech** *speech_file_pathname*

This command performs a recognition trial between the speech contained in the specified file (the unknown) and all templates that are in template memory. There must be at least one template in memory for the DPA process to be initiated.

Each frame of the unknown is compared against all frames of all templates in memory that syntax considers acceptable. Syntax may reject whole templates, or portions of templates due to syntactic restrictions. For a given unknown/template pair, a DPA matrix can be output if both analysis mode is on and the *Analy_tem* environment value is set to the desired template number. This matrix contains detailed information that was used by the DPA process to determine the best matching template. This includes inter-frame distances and indicates the path the DPA process chose as the best one through the matrix (see §1.7 of Data Descriptions for an example).

After all templates have been compared to each frame of the unknown, the unknown phrase is printed along with the best matching choices from among the templates. In continuous speech applications, the unknown is typically a multi-word phrase and the best options are combinations of vocabulary words that proved closest to the unknown. The best matching choice is presented first, along with its score. Following this are the next best choices and scores, one choice to each line, with the better choices appearing first.

3.2.17. **reset_environment** [*variable_name* ...]

The reset_environment command sets environment values back to their initial (default) values. All variables can be reset or just those specified. In either case, those variables that are reset are displayed with their new, initial values.

This command is useful for duplicating the environment for respective experiments in a single run. NOTE: Variables are reset to the values that they have when the program is executed, not necessarily to the values that they had when the first recognition trial was made.

3.2.18. **set_environment** [*variable_name new_value*] ...

This command modifies selected environment values or displays the total environment. The number of arguments determines its processing. If no arguments are present, the CSR environment is displayed with no changes. If arguments are present, they should be in pairs. Each pair specifying a valid environment variable name and a value to set it to. Variable names can be abbreviated to as few characters as are necessary to guarantee uniqueness. It is considered to be a fatal error to reference an unknown or ambiguous variable or to set a variable to a value out of range or of the wrong type.

3.2.19. **signal** *UNIX_signal_name* [off | on]

The signal command causes specified UNIX signals to be ignored or to terminate the program (default setting) upon their receipt. It is a very simple interface to the ***signal*** (2) function that the UNIX kernel makes available. If a given signal is being ignored and it is received by the program, the signal name will be displayed to the user to indicate that the signal was received. This notification can be important since some signals cause system calls (such as reads and writes) to fail, regardless of their effect on the receiving program. For example, the user can specify that interrupts be ignored However, if one is received when a read is outstanding, the read fails and returns an error status.

Although any or all signals can be ignored, it is best to only ignore those signals that are user-generated. Namely, hangup, interrupt and quit. It serves little purpose to ignore a bus error, segmentation violation and others since they indicate very serious program flaws that will cause program termination sooner or later. This command is best used carefully.

3.2.20. **summary** [*one_line_command*]

This summary command causes the current experiment counters to be written out to the experiment results file, following an optional single-line message. If experiment processing is turned off, this command does nothing. Counters written out include the number of phrase (recognition) trials, number of correct options and errors (accompanied by percentages), number of word (sub-recognition) results, correct matches and errors and the number of times the silence template matched. Although the output looks identical to that produced when experiment mode is turned off or when the program terminates normally with experiment processing on, there is one big difference: **summary** does not reset the counters back to zero. The main advantage to using this command is that if multiple speakers are being recognized in an experiment, following each speaker the *accumulated* totals to date can be written out. Notice that since the counters aren't reset, all counts are accumulated from the last time the experiment processing was turned on.

3.2.21. **tty_input**

The tty_input command, when encountered in a command file only, causes the system to audibly prompt the user for keyboard input and continue accepting input from the keyboard until the user enters a control-D or a **quit** command. The system also uses a different prompt (*) to distinguish to the user that this keyboard input has been requested by a command file. This command is useful when the user wishes to check intermediate results during long runs. After the user terminates the keyboard input, processing of the command file continues with the next command. Entering this command from the keyboard elicits a warning message, if the verbose output mode is on.

3.2.22. **unix** [*C_shell_command*]

This command causes a temporary UNIX C shell to be run, temporarily suspending the CSR system for the duration. If no arguments are present, a shell is created and run until the user terminates it with a control-D. Normal shell aliases, search paths and variables are set to those that the user would have when logging in to UNIX. If argument(s) are present, the first one is assumed to be the name of a program to run at the shell level. Second and subsequent arguments are inputs to the program. This version of the shell is historically called a mini-shell since it executes the specified program and immediately exits back to the CSR system. Notice that when arguments are specified to the **unix** command, the user needn't (and shouldn't) enter a control-D to terminate the shell.

3.2.23. **upline_load** *formatted_template_pathname* [*template_#* ...]

The upline_load command allows the user to write templates out of template memory to an external file. The main difference between this command and the **write_template** command is the structure of the output files. This command creates a file that can be read by the **downline_load** command. The file is formatted exactly as template memory is so there is no need for pre-processing of the speech data as there is with the write_template command.

When a single argument is present, it is the pathname of a file to create. If multiple arguments are present, all arguments following the pathname are template numbers to write out. If no numbers are present, all templates are written out in numeric order.

3.2.24. version

The version command causes several lines of information concerning the CSR system to be displayed on the standard output unit. The information includes the program's name, version number, resident directory (of the object) and the last date it was compiled.

3.2.25. write_template *speech_file_pathname* [*template_#* ...]

This command writes templates out of template memory to an external file. All of template memory can be written out or selected templates can be output depending upon whether template numbers are present or not. The output file has the same format as that described in §1.1 of Data Descriptions. Although all of template memory can be written out, it is advised that each template be written to a separate file. The only files that are currently supported that contain multiple speech files are PDP-11/60 archive files (see **archive** command description). Writing the template out does not affect its representation in template memory. Make sure that the current working directory of the program allows writing or specify a rooted pathname as the output file name.

4. Environment Variables

4.1. Variable List

The CSR environment variables control the operation of the system. Most of the environment variables pertain to the recognition process and don't affect any other parts of the system. The environment variables are listed below with a short description, its type (as declared in C), its default value and the range of values that it can be set to. When specifying variable names, only enough characters to guarantee uniqueness are required. Thus, *L*, *Lxm* and *Lxmdim* are all names for the same variable. However, *Amp* is ambiguous since both *Amp_normalize* and *Amp_threshold* begin with the string "Amp".

Environment Variables					
Name	Type	Default	Minimum	Maximum	Comment
Amp_normalize	char	off	off	on	
Amp_threshold	short	200	0	32767	
Analy	char	off	off	on	Read-only
Analy_tem	short	-1	-1	**Templates**	
Beam_factor	float	1.3125	0.	5.	
Beam_threshold	short	8132	0	32767	
Beginpen	short	80	0	32767	
Charmode	char	off	off	on	
Charset	char*	null	*n/a*	*n/a*	
Chrcnt	short	250	0	250	
Chrsz	short	10	0	10	
Dest	char	p	p	v	
Dislim	short	80	0	32767	
Dstcnt	short	80	0	32767	Unused
Dstsz	short	250	0	250	Unused
End_silence	short	8	0	32767	
Exp	char	off	off	on	Read-only
Frame_count	short	500	0	500	
Frame_size	short	20	0	20	
Lxmdim	short	16	0	16	
Maxopt	short	10	0	10	
Min_dst	short	4	0	4	Unused
Mult	short	4	0	32	
Pad	short	5	0	**Frame_count**	
Peak_amp	short	500	0	32767	
Prompt	char*	">"	*n/a*	*n/a*	Limited to 14 bytes
Silshft	short	1	0	32	
Similar	short	1600	0	32767	
Skip	short	2	0	32767	
Source	char	p	p	v	
Templates	short	100	0	100	
Tfrsz	short	20	0	20	
Transform	char*	null	*n/a*	*n/a*	
Verbose	char	on	off	on	
Window	short	10	0	**Frame_count**	
Xformode	char	off	off	on	
Xfrsz	short	10	0	**Frame_size**	
Xscale	short	0	0	32	

Figure X. Summary of CSR Environment Variables

4.2. Environment Variable Descriptions

4.2.1. **Amp_normalize** – Amplitude normalization flag

The amplitude normalization flag controls whether normalization of amplitude is performed on all subsequent inputs of speech data, both templates and unknowns. Amplitude normalization involves replacing the input amplitude with an average amplitude obtained by summing the frame parameters and dividing by the number of parameters. Rounding is used when the number of parameters is not even.

4.2.2. **Amp_threshold** -- Threshold of speech

The amplitude threshold is a value that delimits speech from "silence". It is only used when endpoint detection is being done, usually when live speech is being processed (which is seldom or never). Each input frame's amplitude is compared with the threshold value and if it is less than or equal to the threshold, the frame is assumed to be silence. Conversely, if the amplitude value is greater than the threshold, the speech frame is assumed to be speech.

4.2.3. **Analy** -- Analysis processing mode indicator

This toggle value is read-only and cannot be set using the **set_environment** command. Rather, it indicates in an *off/on* manner the current state of analysis processing. Use the **analysis** command to change the state of analysis mode processing.

4.2.4. **Analy_tem** -- Analysis template number

This value, when set to a template number with analysis processing on, causes the unknown/template DPA matrix and scoring information to be written out to the DPA analysis file. When set to -1 or 0, or when analysis processing is off, does not affect the program or its results.

4.2.5. **Beam_factor** -- Beam search multiplier

The beam_factor defines a window on the partial phrase scores which eliminates some scores from post-processing. After each unknown frame is processed, the beam_factor is multiplied with the best (lowest) score. The resulting product is stored in the *Beam_threshold* environment variable and defines the largest score to process during scoring. Typically, the window defined by the product eliminates 70% of the scores as being too large. The scoring algorithms are very sensitive to this value and even small changes can affect results.

4.2.6. **Beam_threshold** -- Beam search threshold

This value gives an upper-limit on the recognition scores that will be processed following each pass of the unknown frame. It is produced by multiplying the best (lowest) score on a given pass with the *Beam_factor* environment value. It can be set before each recognition trial, however it is reset after processing each unknown frame.

4.2.7. **Beginpen** -- Starting path penalty

This value gives a penalty increment to apply to templates that start paths for each unknown frame. The DPA process and syntax specification may allow templates to begin a path anywhere within the unknown utterance. However, as one proceeds along the unknown, the penalty assigned to templates that begin increases until a path through the DPA matrix ends. This prevents the first few frames from being discarded just because they don't match any of the templates very well. Typically, this value is set to 80 which defines the starting path penalty to be 1 (on unknown frame 0), 81 (frame 1), 161 (frame 2) and so on until a path is completed, at which time the penalty is set to the ending template's score and incremented each unknown frame by *Beginpen*.

APPENDIX B

A CONTINUOUS SPEECH DATA BASE

A CONTINUOUS SPEECH DATA BASE *

B. P. Landell, A. R. Smith, H. M. Koble, and M. L. Alcove

ITT Defense Communications Division
10060 Carroll Canyon Road
San Diego, California 92131

1. INTRODUCTION

The development of continuous speech recognition algorithms requires extensive testing on large data bases from a wide sampling of the speaker population to obtain statistically significant performance data. An attempt to design and record such a data base has been made by the ITT Defense Communications Division. The data base has several useful components including phrase sets generated from progressively larger finite state grammars, a connected digit component, an alphabet spelling component, and a diagnostic rhyme component. This paper describes the content of the data base and the recording facilities and procedures used to collect the speech data.

2. DATA BASE DESCRIPTION

The continuous speech recognition data base designed by the ITT Defense Communications Division is summarized in Table 1. Speech has been recorded from a speaker population consisting of 25 males and 25 females. The spoken material involves training and test utterances from seven data base components which will now be described in detail.

2.1 Airline Sets

The components denoted as Airline Sets 1-4 in Table 1 are interrelated. This portion of the data base has been designed to be

Table 1
DATA BASE SUMMARY

DATA BASE COMPONENT	NUMBER OF SPEAKERS	VOCABULARY SIZE	NUMBER OF VOCABULARY REPETITIONS	CONTINUOUS SPEECH UTTERANCES	NUMBER OF NODES IN GRAMMAR
Airline Set 1	25 male 25 female	53 words	3	50 phrases	21
Airline Set 2	25 male 25 female	100 words	3	50 phrases + 66 phrases for templates	30
Airline Set 3	11 male 11 female	211 words	3	50 phrases	52
Airline Set 4	11 male 11 female	301 words	3	50 phrases	128
Digit Strings	10 male 10 female	10 digits	3	150 digit strings	
Alphabet/ Word Spelling	10 male 10 female	26 letters	6	50 word spellings	
Diagnostic Rhyme -Initial Conson	2 male 2 female	125 words	7		
-Final Conson	2 male 2 female	125 words	7		

* This work was performed under RADC contract number F30602-81-C0155.

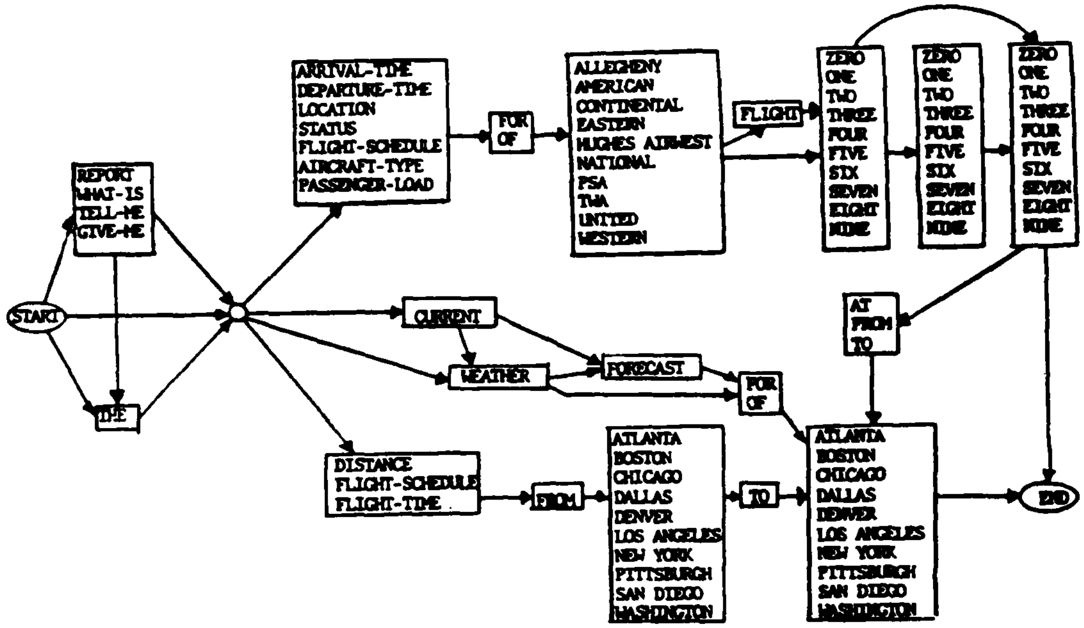


Figure 1. Finite State Grammar For Airline Set 1 (53 word vocabulary, 21 nodes)

representative of limited syntax application areas for speech recognition. The phrases in the Airline Sets are representative of a simplified air travel information retrieval application.

The sentences in Airline Set 1 were generated by the finite state grammar depicted in Figure 1. The grammar contains 21 nodes (including the start and end node) and a 53 word vocabulary. Grammars for Airline Sets 2-4 were then designed by progressively expanding this core finite state grammar to include additional nodes and node connections as well as additional words in the vocabulary. Thus, for example, any phrase generated from the Airline Set 2 grammar can also be generated from the grammar of Airline Sets 3 and 4. However, such a phrase cannot necessarily be generated by the grammar of Airline Set 1. Table 2 gives examples of the types of sentences that are added by Airline Sets 2, 3, and 4.

Table 2
EXAMPLES OF PHRASES FROM AIRLINE SETS 2-4

Airline Set 2:

Report the flight schedule of aircraft charlie yankee four eleven from Boston

Tell me the departure time of Western flight number fifteen.

Airline Set 3:

I want to return from New Orleans to Phoenix on Monday the thirteenth of January

Flight schedule from Hartford Connecticut to Portland Oregon on Friday, April twenty fifth.

Airline Set 4:

I am staying at the Royal Inn hotel

I would like to get some information about my reservation

My telephone number is 929-6685

I will pay with my American Express card.

I would like to report a green coat lost on Allegheny two twelve

Table 3 shows the growth in complexity of the airline grammars by several different measures. The fourth column shows that the maximum sentence length ranges from 11 words to 22 words, while the average sentence length increases from 9.6 words to 18.6. Similarly, in column five, the total number of phrases generated by each grammar ranges by 9 orders of magnitude from 10^8 phrases to 10^{17} phrases. The last column gives an information theoretic measure of complexity. The perplexity* (entropy = $\log_2$ perplexity) measures the average number of word choices at each word position of the language. Thus, for example, a recognition system using Airline Grammar 1 would have to choose between approximately 5.7 words on the average as it identified each new word in the sentence. Without the grammar, but with the same vocabulary, all 53 words would compete at each word position.

Table 3
COMPLEXITY OF AIRLINE GRAMMARS

			Maximum (Average) Words	Number	
Grammar Set	Vocab Size	Nodes	Per Phrase	Of Phrases	Perplexity
Airline 1	53	21	11 (9.6)	$1.0x10^8$	5.74
Airline 2	100	30	14 (12.4)	$3.7x10^{11}$	7.25
Airline 3	211	52	22 (18.6)	$2.1x10^{17}$	7.69
Airline 4	301	126	22 (18.6)	$2.1x10^{17}$	7.69

* Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K., "Perplexity - A Measure of Difficulty of Speech Recognition Tasks", The Journal of the Acoustical Society of America, Vol. 62 S1, 1977, abstract.

According to the perplexity, and to the degree of precision given in the table, Airline Grammar 4 has the same complexity as Airline Grammar 3 even though Set 4 has a larger vocabulary. This is true because the new words of Set 4 are added in new finite state paths which do not significantly increase the total number of phrases (i.e., about 10^7 phrases were added to the 10^{12} phrases in Set 3).

As shown in Table 1, all 50 speakers recorded training and test utterances for Airline Sets 1-2. Training data was collected for each speaker by recording three isolated word repetitions of the 100 word vocabulary associated with Airline Set 2 which included the 53 word subset associated with Airline Set 1. The order of presentation for these words varied with each repetition. Additional training material was obtained by recording 66 phrases generated by the finite state grammar defining Airline Set 2. This set of phrases contains at least two occurrences of each word in the 100 word vocabulary which may be used for extracting word templates.

A total of 22 speakers recorded training and test utterances for Airline Sets 3-4. Training data was collected for each speaker by recording three isolated word repetitions of only the new vocabulary words added by Airline Set 3 (111 words) and Airline Set 4 (90 words). The order of presentation for these additional words varied with each repetition.

The test material for each Airline Set was obtained by recording 50 phrases per speaker. To obtain both a large variety of phrases and common test data across different speakers, the speakers were divided into groups. Each group recorded a different collection of 50 phrases. No more than seven speakers were assigned to the same group. Phrases for each Airline Set were generated randomly from the associated finite state grammar with the following constraints. No phrase was generated twice within a group of 50 phrases (and rarely across any two groups). Also, any phrase generated for Airline Sets 2-4 contained at least one word that was not a member of the vocabulary of the smaller Airline Sets.

2.2 Digit Strings

The fifth component of the data base listed in Table 1 was designed to focus attention on the problem of recognizing strings of connected digits. Template training data was obtained by having each of the 20 speakers record three isolated word repetitions of each digit zero through nine. The order of presentation differed with each repetition. The test data involves a set of 150 digit strings containing 40 three-digit strings, 40 four-digit strings, 50 five-digit strings, and 20 seven-digit strings. The seven-digit strings were presented to the speaker using the pattern XXX-XXXX, thereby giving the appearance of a telephone number. Thus, a total of 670 digits are present in the 150 digit strings.

The list of digit strings have the following additional properties:

1. Every digit (zero through nine) occurs 67 times.
2. Every digit occurs as the first digit 15 times.
3. Every digit occurs as the last digit 15 times.
4. In the 20 strings of seven digits, every digit occurs in the position immediately before the hyphen twice. Every digit occurs in the position immediately after the hyphen twice. Finally, the pair of digits separated by the hyphen are all distinct.

2.3 Alphabet/Word Spellings

The sixth component of the data base shown in Table 1 has been designed to focus attention on the problem of recognizing letters contained in spelled words. Template training data has been obtained from each of the 20 speakers by recording five isolated word repetitions of the alphabet. The order in which the 26 letters were presented varied with each repetition. To obtain test data, each speaker was asked to spell 50 words in a continuous speech fashion. The words were selected from a 4000 word vocabulary with a goal of maximizing the number of unique letter pair combinations in the 50 word subset. These 50 words are shown in Table 4.

Table 4
TEST UTTERANCES FOR WORD SPELLING
COMPONENT OF THE DATA BASE

1. absolution
2. barbarians
3. deployment
4. expounding
5. geographer
6. microscope
7. reassigned
8. childbirth
9. falsehoods
10. brushwork
11. listlessly
12. effectual
13. jackboots
14. overview
15. avoidance
16. adjourning
17. embezzling
18. halfway
19. buckskin
20. immovable
21. ketchup
22. whirlwind
23. attainment
24. suggested
25. cuddly
26. mystique
27. queue
28. thrifty
29. adversary
30. chauvinism
31. dwarf
32. hypocrisy
33. oxygen
34. punchbowl
35. rhythm
36. submerge
37. unifying
38. amazingly
39. anxiously
40. armchair
41. blitzkrieg
42. bluebook
43. breakdown
44. bulky
45. obtains
46. gunpowder
47. relaxation
48. involve
49. knuckle
50. lodgings

2.4 Diagnostic Rhyme Test

The final component of the data base involves a diagnostic rhyme test. The vocabulary for this test was described by J. D. Griffiths*. It consists of 250 words broken into 50 five-word groups. Words within a group differ only in a particular minimal feature. In 25 of the groups, the contrasting element is the final consonant. An example group is: {dig, din, did, dim, dill}. In the remaining 25 groups, the contrasting ele-

* Griffiths, J. D., "Rhyming Minimal Contrasts: A Simplified Diagnostic Articulation Test", The Journal of the Acoustical Society of America, Vol. 42, No. 1, pp. 236-241, July 1967.

ment is the initial consonant. An example group is: {way, may, gay, they, nay}. This component of the data base is useful for analyzing the strengths and weaknesses of a speech recognition system.

A total of eight speakers recorded the diagnostic rhyme test. Four of the speakers recorded the test involving contrasting initial consonants; the other four recorded the test involving contrasting final consonants. The 125 words associated with each half of the test were presented to the speakers in random order. Seven repetitions of the 125 word list were spoken by each speaker with the order of presentation varying for each repetition.

3. EQUIPMENT

The entire data base was recorded in the speech research laboratory at the ITT Defense Communications Division facility in San Diego, California. This laboratory has been carefully designed so that high-quality audio recordings of human speech can be made. A schematic diagram showing the portion of the laboratory which was utilized to record this continuous speech data base is given in Figure 2.

The room labelled "recording room" in the figure contains several features which are conducive to the recording of speech in a quiet atmosphere. The walls are double studded, fiber glass filled, and extend from true floor to true ceiling. A solid core door is used and silencers have been employed in the ventilation unit.

A portion of the room adjacent to the recording room was configured into an operator work station using movable partitions. From this position, the operator had visual contact with the speaker at all times via the double paned glass window between the rooms. The operator's primary function was to guide the speaker through the recording session by controlling both the sequencing and pacing of material displayed on the video monitor in front of the speaker.

3.1 Recording Equipment

Speakers made their speech recordings using a Shure Model SM10 professional headworn microphone. This device is light weight, has a padded headband to minimize user fatigue, and is designed for close talk operation.

Recordings were made using a Technics model RS-1500US tape deck. This unit was placed within easy reach of the operator. All recordings are monaural and were made at a seven and one-half ips tape speed. The unit has a time counter which shows one-half of the actual time for this tape speed.

All recordings were made using one-quarter inch wide, 1200 feet long 3M Scotch 208 audio recording tape which is designed for minimal print through. At seven and one-half ips recording speed, approximately 30 minutes of recording time is possible on a given track in one direction.

The speaker's microphone was connected to a Shure Model M67 professional microphone mixer located in the recording room. A cable was then run from this mixer through the wall and into a line input jack on the rear panel of the recorder at the operator's station. The microphone mixer was required to give adequate gain for weak speakers and to add a synchronization tone at tape startup (described below).

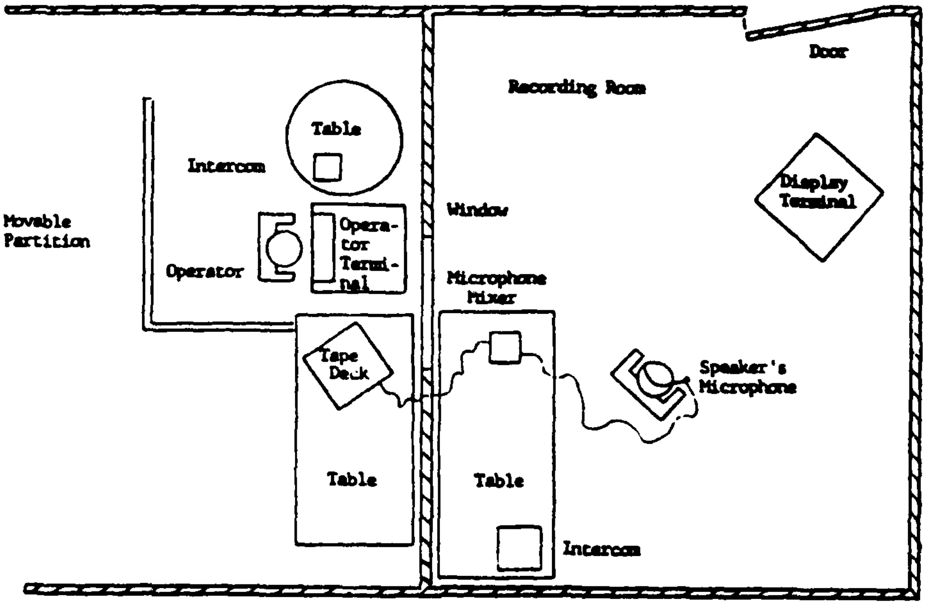


Figure 2. Schematic Diagram of Recording Facilities

3.2 Operator/Speaker Communication Equipment

Several pieces of apparatus facilitated operator/subject communication and automated the display of the material to be recorded.

Two computer terminals were utilized. One terminal was positioned in the recording room for displaying the words and phrases of each recording session to the speaker. The speaker was seated approximately five feet from the terminal. It was found that at distances less than three feet, the electromagnetic radiation emitted from the terminal monitor was picked up as an audible signal by the microphone. The size of the characters in the displayed word or phrase was enlarged for this terminal to minimize speaker eye strain.

The material displayed on the terminal in the recording room was controlled by the operator using a second terminal in the adjacent room. With this terminal, the operator was able to access all data base software files and control the sequencing and pacing of material displayed to the speaker.

To verbally communicate with the speaker, the operator used a push-to-talk intercom. The operator was able to monitor the audio from both the source (i.e the speaker's microphone) and the tape using KOSS Pro4AAA headphones.

4. DATA BASE SOFTWARE

The entire content of the data base as summarized in Table 1 was organized into a set of software text files. These files were stored on a PDP-11/60 computer. With the file structure, the material to be spoken by a given speaker could be defined as a specific sequence of these text files. Moreover, this structure made it possible to intersperse training and test utterances for each component of the data base being recorded by the speaker.

The training and test material for a given recording session was presented to the speaker on the video terminal using prompting software specially created for this project. In order to avoid a list reading style, this software displayed one utterance (word or phrase) at a time from the specified text file to the speaker. It also displayed the utterance on the terminal at the operator's station, along with an index number.

The prompting software enabled the operator via keyboard commands to control the material spoken by the subject. After the speaker completed his or her response to the displayed word or phrase, the operator could (a) enter a "continue" command which would cause the prompting software to display the next utterance in the text file, (b) enter a "repeat" command to indicate that the speaker had corrected an error made in uttering the word or phrase, or (c) enter a command to have a specified utterance from the file displayed to the speaker. Option "c" was primarily utilized by the operator to redisplay the word or phrase just uttered by the speaker if an error had been made which was not self-corrected by the speaker.

The prompting software had one additional feature. As the session proceeded, the prompting software created a recording history file. This file contained a record of each utterance spoken by the speaker, and the relative time (in tape counter units) at which the speaker was prompted. All repeat commands were also included in the history file along with all line redisplays commanded by the operator. Each time the audio tape was stopped during the recording session, the operator entered the current counter reading from the tape recorder into the history file. Then, upon restarting the tape, a tone sequence was generated by the software and recorded on the tape prior to display of the first utterance in the text file. This tone sequence is useful during tape playback, audio tape editing, or digitization

The relative prompt times were generated by the computer clock and were initialized relative to the start-up of the tape recorder and corresponding tone sequence. Thus, a position in the data immediately preceding each spoken utterance can be accurately located.

5. DATA BASE COLLECTION

A typical recording session lasted between one and one-half and two hours. Most speakers had very limited information regarding the nature of the activity they were about to perform.

Prior to entering the recording room, each speaker completed a questionnaire in which the following information was solicited: sex, age, height, weight, place of birth, residences since birth(city, state, country, dates resided therein), educational history (name of school, location, dates, degree, major field), employment history (occupation, dates), military experience, languages other than English spoken fluently (including whether learned as a child, teenager, or adult and the source of this knowledge). In addition, the questionnaire asked the speaker if he or she had any unusual characteristics in their conversational speech patterns and to describe any previous experience they may have had as a subject in a speech recognition experiment. While the speaker was filling out this questionnaire, the operator was checking the recording equipment and conditions in the laboratory to ensure that the configuration was proper for the upcoming recording session.

Once the questionnaire had been completed, the speaker was escorted into the recording room and seated. The operator then provided the speaker with specific information regarding the recording session. Speaking from a set of notes, the operator described the content of the session and the mechanism which would be used to present material to the speaker for recording. The operator advised the subjects to speak as naturally as possible. The speakers were told that if they caught themselves saying something incorrectly, they were to say the word "correction" and then repeat the word or phrase being displayed. They were also advised that the operator would redisplay the same word or phrase for re-utterance if the operator felt the speaker had misspoken the word or phrase. The operator also discussed any idiosyncrasies which might be present in the content of a given component of the data base. For example, if the subjects were recording the digit string component, they were cautioned to say "zero" instead of "oh" if the symbol "0" appeared in the digit string sequence. If the subjects were recording one of the diagnostic rhyme components, they were told that a few of the words had commonly used multiple pronunciations. Since only one of these pronunciations was acceptable for this test, the subjects were advised that the construction "sounds like ..." followed by a rhyming word would appear to assist them in deducing the correct pronunciation. During this briefing, the subjects were free to ask questions and make any other remarks which would add to their understanding of the task at hand.

After completing this briefing, the operator placed the headset on the speaker and properly positioned the Shure SM10 unidirectional microphone. The speaker was cautioned about touching this apparatus or making any head movements which might alter its position. Then, the operator left the recording room and returned to the operator station in the adjacent room of the laboratory.

Next, the operator performed a microphone test prior to initiating recording. In this test, the speakers were shown a series of four phrases and asked to say them using their natural speaking voice. The operator adjusted the recording level for the given speaker by monitoring the VU meter reading on the tape recorder. The level was considered to be properly adjusted when the peaks of the recording level were barely into the positive db range and the average recording level was roughly -3 to -5 db. For some speakers, this microphone test had to be repeated two or more times before the operator was satisfied with the recording level calibration. Occasionally, the operator made additional microphone tests during the session, if the initial calibration setting became unacceptable due to a change in the level of a speaker's speech.

Once the microphone test was completed, the recorder was started and the subject recorded a preamble message which identified the date of the recording, the speaker by name, the speaker number, and the session content. The recorder was then stopped long enough to remind the speaker about the content and any idiosyncrasies of the first segment of speech to be recorded. The session then proceeded with the operator controlling the pace of the the pre-defined sequence of material which the speaker recorded. After each vocabulary repetition or test phrase section was completed, the recorder was again briefly stopped to remind the speaker about the content of the segment which followed.

A total of 50 speakers were recorded; 25 males and 25 females. In response to contractual commitments, the vast majority of these individuals had little or no previous experience as subjects in experiments involving voice recording for speech recognition purposes. The subject population was drawn from two sources. Roughly one-half of the speakers were employees of the ITT Defense Communications Division San Diego facility at the time the recordings were made. The other half of the speaker population was affiliated with a temporary employment agency located in the San Diego area.

Male subjects ranged in age from 21-51; the median age was 29. Female subjects ranged in age from 18-58; the median age was also 29. Since the San Diego population represents a melting pot of people from throughout the United States, the subject population represents a broad mix of native American dialects.

The 50 subjects were divided into four type classifications. Regardless of the classification category, each speaker first recorded training and test material from Airline Sets 1-2.

Then, depending upon the classification type, the speaker recorded training and test utterances from other components of the data base as identified in Table 1. Type A speakers recorded material from Airline Sets 3-4; Type B speakers recorded material from the Diagnostic Rhyme test involving contrasting initial consonants; Type C speakers recorded material from the Diagnostic Rhyme test involving contrasting final consonants; and Type D speakers recorded material from the digit string and alphabet/word spelling component of the data base.

The vast majority of the speakers performed the Airline Set 1-2 task in 25 to 30 minutes of actual recording time. Each speaker was given a 20 minute break before beginning the second half of the task. The recording time for the second session ranged from approximately 18 minutes to 37 minutes, depending upon the speakers and the classification type.

6. SUMMARY

This speech data base plays an integral part in the ITT Defense Communications Division's continuing effort to develop efficient, effective continuous speech recognition algorithms. The data is restricted to speech from cooperative speakers recorded under quiet conditions. Within the boundaries of this environment, however, the scope of the data base is quite broad. Speech has been recorded from 50 speakers and involves training and test utterances from seven different components. Four of these are representative of limited syntax application areas for speech recognition. The others involve connected digits, use of the alphabet in a continuous speech manner to spell words, and a diagnostic rhyme component. Special care was taken to produce quality recordings. The spoken material was displayed on a video monitor. Prompting and pacing of this material was controlled by an operator positioned in an adjacent room. Although the speaker population was drawn from only two sources - ITT Defense Communications Division employees and personnel from a temporary employment agency - a good demographic mixture with respect to age, dialect, and physical size, was obtained. The recordings also contain typical non-speech sounds, such as lip smacking, tongue clicking and breathing. In addition, speaking styles varied from a rapid and heavily coarticulated, to slow and deliberate. The data base thus provides a valuable tool for developing and testing speech recognition systems.

APPENDIX C

PERFORMANCE TEST RESULTS

BY INDIVIDUAL SPEAKER

INTERPRETATION GUIDE FOR EXPERIMENT SUMMARIES

"*speaker nn*" - nn is speaker number followed by sex, age, and highest level of education

"ALL WORDS" - recognition results for all vocabulary words.

"SEMANTIC" - recognition results excluding "of,for,the"

"CORRECT" - number of words or phrases correctly identified.

"OPTION n" - number of phrases for which the CSR system's nth candidate phrase was correct.

"ERRORS" - number of phrases for which none of CSR system's candidate phrases were correct.

```
EXPERIMENT SUMMARY
speaker 01: male, 29, high school
                ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  18   36.0%      38   76.0%
  OPTION #2 =  12   24.0%       5   10.0%
  OPTION #3 =   1    2.0%       1    2.0%
  OPTION #4 =   1    2.0%       1    2.0%
  OPTION #5 =   2    4.0%       0    0.0%
     ERRORS =  16   32.0%       5   10.0%

 WORD TRIALS =  353            286
    CORRECT =  308             273
 INSERTIONS =    0               1
  DELETIONS =   12               2
  WORD RATE =       87.3%          95.1%

EXPERIMENT SUMMARY
speaker 02: female, 28, high school
                ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  39   78.0%      46   92.0%
  OPTION #2 =   7   14.0%       1    2.0%
  OPTION #3 =   0    0.0%       0    0.0%
  OPTION #4 =   0    0.0%       0    0.0%
  OPTION #5 =   1    2.0%       0    0.0%
     ERRORS =   3    6.0%       3    6.0%

 WORD TRIALS =  358            291
    CORRECT =  346             284
 INSERTIONS =    5               3
  DELETIONS =    1               1
  WORD RATE =       95.3%          96.6%

EXPERIMENT SUMMARY
speaker 03: female, 26, junior college
                ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  12   24.0%      27   54.0%
  OPTION #2 =  17   34.0%       3    6.0%
  OPTION #3 =   1    2.0%       2    4.0%
  OPTION #4 =   1    2.0%       0    0.0%
  OPTION #5 =   1    2.0%       1    2.0%
     ERRORS =  18   36.0%      17   34.0%

 WORD TRIALS =  392            313
    CORRECT =  339             280
 INSERTIONS =    2               0
  DELETIONS =    4               2
  WORD RATE =       86.0%          89.5%
```

```
EXPERIMENT SUMMARY
speaker 04: male, 22, junior college
              ALL WORDS           SEMANTIC
PHRASE TRIALS =  50
     CORRECT = 30  60.0%      42  84.0%
   OPTION #2 =  6  12.0%       4   8.0%
   OPTION #3 =  2   4.0%       1   2.0%
   OPTION #4 =  0   0.0%       0   0.0%
   OPTION #5 =  1   2.0%       1   2.0%
      ERRORS = 11  22.0%       2   4.0%

  WORD TRIALS =  375          299
     CORRECT = 356            293
   INSERTIONS =   3             2
    DELETIONS =   9             0
    WORD RATE =     94.2%        97.3%


EXPERIMENT SUMMARY
speaker 05: male, 30, bachelors
              ALL WORDS           SEMANTIC
PHRASE TRIALS =  50
     CORRECT = 37  74.0%      47  94.0%
   OPTION #2 =  5  10.0%       0   0.0%
   OPTION #3 =  0   0.0%       0   0.0%
   OPTION #4 =  1   2.0%       1   2.0%
   OPTION #5 =  0   0.0%       0   0.0%
      ERRORS =  7  14.0%       2   4.0%

  WORD TRIALS =  364          290
     CORRECT = 352            289
   INSERTIONS =   2             2
    DELETIONS =   7             0
    WORD RATE =     96.2%        99.0%


EXPERIMENT SUMMARY
speaker 06: female, 30, phd.
              ALL WORDS           SEMANTIC
PHRASE TRIALS =  50
     CORRECT = 27  54.0%      39  78.0%
   OPTION #2 = 10  20.0%       3   6.0%
   OPTION #3 =  4   8.0%       2   4.0%
   OPTION #4 =  2   4.0%       1   2.0%
   OPTION #5 =  0   0.0%       0   0.0%
      ERRORS =  7  14.0%       5  10.0%

  WORD TRIALS =  353          278
     CORRECT = 325            267
   INSERTIONS =   3             3
    DELETIONS =   6             0
    WORD RATE =     91.3%        95.0%
```

```
EXPERIMENT SUMMARY
speaker 07: female, 28, high school
               ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  40   80.0%      44   88.0%
   OPTION #2 =   5   10.0%       3    6.0%
   OPTION #3 =   1    2.0%       0    0.0%
   OPTION #4 =   0    0.0%       0    0.0%
   OPTION #5 =   0    0.0%       0    0.0%
      ERRORS =   4    8.0%       3    6.0%

  WORD TRIALS =  380            307
     CORRECT =  368             299
  INSERTIONS =    0               0
   DELETIONS =    3               1
   WORD RATE =       96.8%          97.4%


EXPERIMENT SUMMARY
speaker 08: male, 34, phd.
               ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  29   58.0%      41   82.0%
   OPTION #2 =   8   16.0%       5   10.0%
   OPTION #3 =   2    4.0%       1    2.0%
   OPTION #4 =   1    2.0%       0    0.0%
   OPTION #5 =   1    2.0%       1    2.0%
      ERRORS =   9   18.0%       2    4.0%

  WORD TRIALS =  388            311
     CORRECT =  373             304
  INSERTIONS =    9               2
   DELETIONS =    6               0
   WORD RATE =       94.0%          97.1%


EXPERIMENT SUMMARY
speaker 09: male, 30, masters
               ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  27   54.0%      40   80.0%
   OPTION #2 =  14   28.0%       7   14.0%
   OPTION #3 =   1    2.0%       0    0.0%
   OPTION #4 =   3    6.0%       1    2.0%
   OPTION #5 =   1    2.0%       1    2.0%
      ERRORS =   4    8.0%       1    2.0%

  WORD TRIALS =  375            299
     CORRECT =  353             290
  INSERTIONS =    1               1
   DELETIONS =   10               0
   WORD RATE =       93.9%          96.7%
```

```
EXPERIMENT SUMMARY
speaker 10: female, 35, bachelors
                 ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
      CORRECT =  34   68.0%       43   86.0%
    OPTION #2 =   8   16.0%        4    8.0%
    OPTION #3 =   1    2.0%        0    0.0%
    OPTION #4 =   1    2.0%        0    0.0%
    OPTION #5 =   0    0.0%        0    0.0%
       ERRORS =   6   12.0%        3    6.0%

   WORD TRIALS =  361              288
       CORRECT =  342              278
    INSERTIONS =    2                0
     DELETIONS =   10                4
     WORD RATE =      94.2%          96.5%
```

```
EXPERIMENT SUMMARY
speaker 11: female, 23, junior college
                 ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
      CORRECT =  35   70.0%       49   98.0%
    OPTION #2 =   4    8.0%        0    0.0%
    OPTION #3 =   1    2.0%        0    0.0%
    OPTION #4 =   2    4.0%        0    0.0%
    OPTION #5 =   1    2.0%        0    0.0%
       ERRORS =   7   14.0%        1    2.0%

   WORD TRIALS =  353              278
       CORRECT =  333              276
    INSERTIONS =    1                0
     DELETIONS =    8                0
     WORD RATE =      94.1%          99.3%
```

```
EXPERIMENT SUMMARY
speaker 12: male, 32, junior college
                 ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
      CORRECT =  33   66.0%       42   84.0%
    OPTION #2 =   5   10.0%        3    6.0%
    OPTION #3 =   2    4.0%        0    0.0%
    OPTION #4 =   1    2.0%        0    0.0%
    OPTION #5 =   1    2.0%        1    2.0%
       ERRORS =   8   16.0%        4    8.0%

   WORD TRIALS =  378              306
       CORRECT =  352              295
    INSERTIONS =    2                2
     DELETIONS =    9                0
     WORD RATE =      92.6%          95.8%
```

```
EXPERIMENT SUMMARY
speaker 13: female, 43, bachelors
                ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
      CORRECT = 40   80.0%     46  92.0%
    OPTION #2 =  5   10.0%      3   6.0%
    OPTION #3 =  1    2.0%      0   0.0%
    OPTION #4 =  0    0.0%      0   0.0%
    OPTION #5 =  0    0.0%      0   0.0%
       ERRORS =  4    8.0%      1   2.0%

  WORD TRIALS = 382            310
      CORRECT = 368            304
   INSERTIONS =   0              0
    DELETIONS =   7              1
    WORD RATE =      96.3%        98.1%
```

```
EXPERIMENT SUMMARY
speaker 14: male, 22, high school
                ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
      CORRECT = 26   52.0%     40  80.0%
    OPTION #2 =  4    8.0%      4   8.0%
    OPTION #3 =  1    2.0%      0   0.0%
    OPTION #4 =  1    2.0%      0   0.0%
    OPTION #5 =  1    2.0%      0   0.0%
       ERRORS = 17   34.0%      6  12.0%

  WORD TRIALS = 381            307
      CORRECT = 348            291
   INSERTIONS =   2              1
    DELETIONS =  13              3
    WORD RATE =      90.9%        94.5%
```

```
EXPERIMENT SUMMARY
speaker 15: male, 29, bachelors
                ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
      CORRECT = 32   64.0%     37  74.0%
    OPTION #2 =  6   12.0%      6  12.0%
    OPTION #3 =  4    8.0%      2   4.0%
    OPTION #4 =  1    2.0%      1   2.0%
    OPTION #5 =  0    0.0%      0   0.0%
       ERRORS =  7   14.0%      4   8.0%

  WORD TRIALS = 359            286
      CORRECT = 334            270
   INSERTIONS =   2              2
    DELETIONS =   6              1
    WORD RATE =      92.5%        93.8%
```

```
EXPERIMENT SUMMARY
speaker 16: male, 48, bachelors
              ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  19   38.0%     41   82.0%
  OPTION #2 =   6   12.0%      1    2.0%
  OPTION #3 =   3    6.0%      1    2.0%
  OPTION #4 =   1    2.0%      0    0.0%
  OPTION #5 =   0    0.0%      0    0.0%
     ERRORS =  21   42.0%      7   14.0%

 WORD TRIALS =  355           279
     CORRECT =  300           256
  INSERTIONS =    4             3
   DELETIONS =   14             3
   WORD RATE =      83.6%        90.8%
```

```
EXPERIMENT SUMMARY
speaker 17: male, 42, masters
              ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  37   74.0%     42   84.0%
  OPTION #2 =   5   10.0%      4    8.0%
  OPTION #3 =   2    4.0%      0    0.0%
  OPTION #4 =   1    2.0%      0    0.0%
  OPTION #5 =   0    0.0%      0    0.0%
     ERRORS =   5   10.0%      4    8.0%

 WORD TRIALS =  378           306
     CORRECT =  365           299
  INSERTIONS =    2             1
   DELETIONS =    3             0
   WORD RATE =      96.1%        97.4%
```

```
EXPERIMENT SUMMARY
speaker 18: male, 33, masters
              ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  21   42.0%     35   70.0%
  OPTION #2 =   7   14.0%      3    6.0%
  OPTION #3 =   3    6.0%      2    4.0%
  OPTION #4 =   2    4.0%      2    4.0%
  OPTION #5 =   0    0.0%      0    0.0%
     ERRORS =  17   34.0%      8   16.0%

 WORD TRIALS =  386           311
     CORRECT =  349           288
  INSERTIONS =   11             6
   DELETIONS =   12             4
   WORD RATE =      87.5%        90.9%
```

```
EXPERIMENT SUMMARY
speaker 19: male, 34, bachelors
                ALL WORDS           SEMANTIC
PHRASE TRIALS =  50
      CORRECT =  39   78.0%      44  88.0%
    OPTION #2 =   9   18.0%       4   8.0%
    OPTION #3 =   1    2.0%       1   2.0%
    OPTION #4 =   0    0.0%       0   0.0%
    OPTION #5 =   0    0.0%       0   0.0%
       ERRORS =  1    2.0%        1   2.0%

 WORD TRIALS =  375              299
     CORRECT =  366              295
  INSERTIONS =    3                3
   DELETIONS =    3                0
   WORD RATE =      96.8%          97.7%
```

```
EXPERIMENT SUMMARY
speaker 20: male, 29, bachelors
                ALL WORDS           SEMANTIC
PHRASE TRIALS =  50
      CORRECT =  37   74.0%      42  84.0%
    OPTION #2 =   7   14.0%       5  10.0%
    OPTION #3 =   0    0.0%       0   0.0%
    OPTION #4 =   1    2.0%       0   0.0%
    OPTION #5 =   0    0.0%       0   0.0%
       ERRORS =   5   10.0%       3   6.0%

 WORD TRIALS =  364              290
     CORRECT =  347              280
  INSERTIONS =    1                0
   DELETIONS =    5                2
   WORD RATE =      95.1%          96.6%
```

```
EXPERIMENT SUMMARY
speaker 21: male, 29, bachelors
                ALL WORDS           SEMANTIC
PHRASE TRIALS =  50
      CORRECT =  37   74.0%      44  88.0%
    OPTION #2 =   6   12.0%       4   8.0%
    OPTION #3 =   1    2.0%       0   0.0%
    OPTION #4 =   2    4.0%       0   0.0%
    OPTION #5 =   0    0.0%       0   0.0%
       ERRORS =  4    8.0%        2   4.0%

 WORD TRIALS =  353              278
     CORRECT =  338              270
  INSERTIONS =    2                0
   DELETIONS =    5                0
   WORD RATE =      95.2%          97.1%
```

```
EXPERIMENT SUMMARY
speaker 22: male, 23, bachelors
               ALL WORDS         SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  38   76.0%      41   82.0%
  OPTION #2 =   6   12.0%       4    8.0%
  OPTION #3 =   0    0.0%       0    0.0%
  OPTION #4 =   1    2.0%       1    2.0%
  OPTION #5 =   0    0.0%       0    0.0%
     ERRORS =   5   10.0%       4    8.0%

  WORD TRIALS =  378           306
     CORRECT =  366            298
  INSERTIONS =    1              2
   DELETIONS =    2              0
   WORD RATE =      96.6%         96.8%
```

```
EXPERIMENT SUMMARY
speaker 23: male, 33, masters
               ALL WORDS         SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  26   52.0%      40   80.0%
  OPTION #2 =   9   18.0%       4    8.0%
  OPTION #3 =   3    6.0%       0    0.0%
  OPTION #4 =   0    0.0%       1    2.0%
  OPTION #5 =   0    0.0%       0    0.0%
     ERRORS =  12   24.0%       5   10.0%

  WORD TRIALS =  391           314
     CORRECT =  363            300
  INSERTIONS =    4              2
   DELETIONS =   13              3
   WORD RATE =      91.9%         94.9%
```

```
EXPERIMENT SUMMARY
speaker 24: female, 36, phd.
               ALL WORDS         SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  35   70.0%      43   86.0%
  OPTION #2 =   7   14.0%       4    8.0%
  OPTION #3 =   1    2.0%       1    2.0%
  OPTION #4 =   0    0.0%       0    0.0%
  OPTION #5 =   1    2.0%       0    0.0%
     ERRORS =   6   12.0%       2    4.0%

  WORD TRIALS =  380           302
     CORRECT =  362            294
  INSERTIONS =    1              1
   DELETIONS =    8              1
   WORD RATE =      95.0%         97.0%
```

```
EXPERIMENT SUMMARY
speaker 25: female, 43, high school
                 ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
     CORRECT = 35   70.0%      41   82.0%
   OPTION #2 =  7   14.0%       3    6.0%
   OPTION #3 =  0    0.0%       0    0.0%
   OPTION #4 =  0    0.0%       0    0.0%
   OPTION #5 =  0    0.0%       1    2.0%
      ERRORS =  8   16.0%       5   10.0%

 WORD TRIALS =  373             301
    CORRECT =  356              289
  INSERTIONS =   3                1
   DELETIONS =   6                1
   WORD RATE =     94.7%          95.7%
```

```
EXPERIMENT SUMMARY
speaker 26: male, 32, masters
                 ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
     CORRECT = 43   86.0%      47   94.0%
   OPTION #2 =  2    4.0%       2    4.0%
   OPTION #3 =  1    2.0%       0    0.0%
   OPTION #4 =  0    0.0%       0    0.0%
   OPTION #5 =  0    0.0%       0    0.0%
      ERRORS =  4    8.0%       1    2.0%

 WORD TRIALS =  352             277
    CORRECT =  344              273
  INSERTIONS =   0                0
   DELETIONS =   4                1
   WORD RATE =     97.7%          98.6%
```

```
EXPERIMENT SUMMARY
speaker 27: female, 28, masters
                 ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
     CORRECT = 39   78.0%      45   90.0%
   OPTION #2 =  6   12.0%       2    4.0%
   OPTION #3 =  1    2.0%       0    0.0%
   OPTION #4 =  1    2.0%       1    2.0%
   OPTION #5 =  0    0.0%       0    0.0%
      ERRORS =  3    6.0%       2    4.0%

 WORD TRIALS =  378             306
    CORRECT =  365              300
  INSERTIONS =   1                0
   DELETIONS =   5                1
   WORD RATE =     96.3%          98.0%
```

```
EXPERIMENT SUMMARY
speaker 28: female, 58, bachelors
                   ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
       CORRECT =  40   80.0%       47   94.0%
     OPTION #2 =   5   10.0%        2    4.0%
     OPTION #3 =   1    2.0%        0    0.0%
     OPTION #4 =   0    0.0%        0    0.0%
     OPTION #5 =   0    0.0%        0    0.0%
        ERRORS =   4    8.0%        1    2.0%

   WORD TRIALS =  342              277
       CORRECT =  333              274
    INSERTIONS =    1                0
     DELETIONS =    1                0
     WORD RATE =       97.1%           98.9%
```

```
EXPERIMENT SUMMARY
speaker 29: male, 34, junior college
                   ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
       CORRECT =  33   66.0%       43   86.0%
     OPTION #2 =   7   14.0%        4    8.0%
     OPTION #3 =   2    4.0%        1    2.0%
     OPTION #4 =   1    2.0%        0    0.0%
     OPTION #5 =   1    2.0%        1    2.0%
        ERRORS =   6   12.0%        1    2.0%

   WORD TRIALS =  375              299
       CORRECT =  355              292
    INSERTIONS =    0                0
     DELETIONS =    9                0
     WORD RATE =       94.7%           97.7%
```

```
EXPERIMENT SUMMARY
speaker 30: male, 26, bachelors
                   ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
       CORRECT =  25   50.0%       39   78.0%
     OPTION #2 =   6   12.0%        5   10.0%
     OPTION #3 =   1    2.0%        2    4.0%
     OPTION #4 =   1    2.0%        0    0.0%
     OPTION #5 =   1    2.0%        0    0.0%
        ERRORS =  16   32.0%        4    8.0%

   WORD TRIALS =  364              290
       CORRECT =  338              279
    INSERTIONS =    7                1
     DELETIONS =   10                1
     WORD RATE =       91.1%           95.9%
```

```
EXPERIMENT SUMMARY
speaker 31: male, 27, junior college
                ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
      CORRECT = 37  74.0%     44  88.0%
    OPTION #2 = 10  20.0%      5  10.0%
    OPTION #3 =  1   2.0%      0   0.0%
    OPTION #4 =  1   2.0%      0   0.0%
    OPTION #5 =  0   0.0%      0   0.0%
       ERRORS =  1   2.0%      1   2.0%

   WORD TRIALS = 357           291
     CORRECT = 348             288
   INSERTIONS =   4              3
    DELETIONS =   3              0
    WORD RATE =      96.4%        98.0%

EXPERIMENT SUMMARY
speaker 32: female, 24, junior college
                ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
      CORRECT = 39  78.0%     47  94.0%
    OPTION #2 =  6  12.0%      1   2.0%
    OPTION #3 =  1   2.0%      0   0.0%
    OPTION #4 =  0   0.0%      0   0.0%
    OPTION #5 =  1   2.0%      0   0.0%
       ERRORS =  3   6.0%      2   4.0%

   WORD TRIALS = 377           303
     CORRECT = 364             300
   INSERTIONS =   0              0
    DELETIONS =   7              0
    WORD RATE =      96.6%        99.0%

EXPERIMENT SUMMARY
speaker 33: female, 51, junior college
                ALL WORDS        SEMANTIC
PHRASE TRIALS =  50
      CORRECT = 37  74.0%     45  90.0%
    OPTION #2 =  5  10.0%      2   4.0%
    OPTION #3 =  2   4.0%      2   4.0%
    OPTION #4 =  1   2.0%      0   0.0%
    OPTION #5 =  0   0.0%      0   0.0%
       ERRORS =  5  10.0%      1   2.0%

   WORD TRIALS = 375           299
     CORRECT = 364             295
   INSERTIONS =   2              1
    DELETIONS =   5              0
    WORD RATE =      96.6%        98.3%
```

```
EXPERIMENT SUMMARY
speaker 34: female, 48, high school
               ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  48   96.0%      50  100.0%
  OPTION #2 =  2    4.0%       0    0.0%
  OPTION #3 =  0    0.0%       0    0.0%
  OPTION #4 =  0    0.0%       0    0.0%
  OPTION #5 =  0    0.0%       0    0.0%
     ERRORS =  0    0.0%       0    0.0%

 WORD TRIALS =  358             291
    CORRECT =  356              291
  INSERTIONS =   0               0
   DELETIONS =   1               0
   WORD RATE =      99.4%        100.0%


EXPERIMENT SUMMARY
speaker 35: female, 24, junior college
               ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  33   66.0%      38   76.0%
  OPTION #2 =  5   10.0%       2    4.0%
  OPTION #3 =  3    6.0%       2    4.0%
  OPTION #4 =  1    2.0%       0    0.0%
  OPTION #5 =  1    2.0%       1    2.0%
     ERRORS =  7   14.0%       7   14.0%

 WORD TRIALS =  375             303
    CORRECT =  357              289
  INSERTIONS =   5               3
   DELETIONS =   3               1
   WORD RATE =      93.9%         94.4%


EXPERIMENT SUMMARY
speaker 36: female, 29, bachelors
               ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  34   68.0%      45   90.0%
  OPTION #2 =  9   18.0%       3    6.0%
  OPTION #3 =  2    4.0%       1    2.0%
  OPTION #4 =  1    2.0%       0    0.0%
  OPTION #5 =  0    0.0%       0    0.0%
     ERRORS =  4    8.0%       1    2.0%

 WORD TRIALS =  340             279
    CORRECT =  325              274
  INSERTIONS =   1               0
   DELETIONS =   3               0
   WORD RATE =      95.3%         98.2%
```

```
EXPERIMENT SUMMARY
speaker 37: female, 28, masters
            ALL WORDS            SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  16   32.0%      23   46.0%
   OPTION #2 =   6   12.0%       3    6.0%
   OPTION #3 =   2    4.0%       1    2.0%
   OPTION #4 =   0    0.0%       0    0.0%
   OPTION #5 =   1    2.0%       1    2.0%
      ERRORS =  25   50.0%      22   44.0%

  WORD TRIALS =  368            307
     CORRECT =  294             252
   INSERTIONS =   5               1
    DELETIONS =  24              17
    WORD RATE =      78.8%         81.8%
```

```
EXPERIMENT SUMMARY
speaker 38: male, 24, bachelors
            ALL WORDS            SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  26   52.0%      39   78.0%
   OPTION #2 =   9   18.0%       4    8.0%
   OPTION #3 =   0    0.0%       0    0.0%
   OPTION #4 =   1    2.0%       0    0.0%
   OPTION #5 =   1    2.0%       1    2.0%
      ERRORS =  13   26.0%       6   12.0%

  WORD TRIALS =  367            305
     CORRECT =  338             292
   INSERTIONS =   5               3
    DELETIONS =  14               2
    WORD RATE =      90.9%         94.8%
```

```
EXPERIMENT SUMMARY
speaker 39: female, 28, junior college
            ALL WORDS            SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  40   80.0%      44   88.0%
   OPTION #2 =   4    8.0%       1    2.0%
   OPTION #3 =   4    8.0%       3    6.0%
   OPTION #4 =   0    0.0%       0    0.0%
   OPTION #5 =   0    0.0%       0    0.0%
      ERRORS =   2    4.0%       2    4.0%

  WORD TRIALS =  367            307
     CORRECT =  355             299
   INSERTIONS =   0               0
    DELETIONS =   2               0
    WORD RATE =      96.7%         97.4%
```

```
EXPERIMENT SUMMARY
speaker 40: male, 23, bachelors
                  ALL WORDS              SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  36   72.0%         43   86.0%
   OPTION #2 =   3    6.0%          5   10.0%
   OPTION #3 =   0    0.0%          0    0.0%
   OPTION #4 =   0    0.0%          0    0.0%
   OPTION #5 =   0    0.0%          0    0.0%
      ERRORS =  11   22.0%          2    4.0%

 WORD TRIALS =  379                307
     CORRECT =  363                300
  INSERTIONS =    0                  0
   DELETIONS =    8                  0
   WORD RATE =       95.8%             97.7%
```

```
EXPERIMENT SUMMARY
speaker 41: female, 18, high school
                  ALL WORDS              SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  34   68.0%         41   82.0%
   OPTION #2 =   2    4.0%          2    4.0%
   OPTION #3 =   1    2.0%          2    4.0%
   OPTION #4 =   2    4.0%          1    2.0%
   OPTION #5 =   1    2.0%          0    0.0%
      ERRORS =  10   20.0%          4    8.0%

 WORD TRIALS =  357                288
     CORRECT =  338                277
  INSERTIONS =    3                  1
   DELETIONS =    3                  1
   WORD RATE =       93.9%             95.8%
```

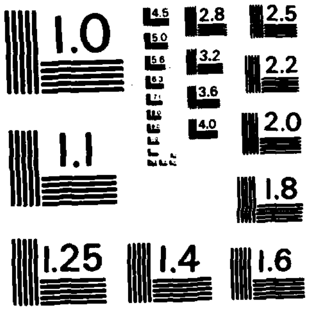
```
EXPERIMENT SUMMARY
speaker 42: female, 28, bachelors
                  ALL WORDS              SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  29   58.0%         39   78.0%
   OPTION #2 =   2    4.0%          2    4.0%
   OPTION #3 =   5   10.0%          3    6.0%
   OPTION #4 =   1    2.0%          1    2.0%
   OPTION #5 =   2    4.0%          0    0.0%
      ERRORS =  11   22.0%          5   10.0%

 WORD TRIALS =  374                302
     CORRECT =  352                290
  INSERTIONS =   11                  6
   DELETIONS =    4                  0
   WORD RATE =       91.4%             94.2%
```

AD-A131 048   LIMITED CONNECTED SPEECH EXPERIMENT(U) ITT DEFENSE
COMMUNICATIONS DIV NUTLEY N J   P B LANDELL MAR 83
RADC-TR-82-316 F30602-81-C-0155

2/2

UNCLASSIFIED   F/G 17/2   NL

END
DATE
FILMED
6 83
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

```
EXPERIMENT SUMMARY
speaker 43: male, 21, junior college
          ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  32  64.0%     40  80.0%
  OPTION #2 =   4   8.0%      4   8.0%
  OPTION #3 =   1   2.0%      0   0.0%
  OPTION #4 =   1   2.0%      0   0.0%
  OPTION #5 =   0   0.0%      1   2.0%
     ERRORS =  12  24.0%      5  10.0%

 WORD TRIALS =  386          313
    CORRECT =  362           301
 INSERTIONS =    2             1
  DELETIONS =   13             1
  WORD RATE =      93.3%        95.9%
```

```
EXPERIMENT SUMMARY
speaker 44: female, 36, junior college
          ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  37  74.0%     41  82.0%
  OPTION #2 =   4   8.0%      1   2.0%
  OPTION #3 =   3   6.0%      2   4.0%
  OPTION #4 =   0   0.0%      0   0.0%
  OPTION #5 =   1   2.0%      1   2.0%
     ERRORS =   5  10.0%      5  10.0%

 WORD TRIALS =  386          314
    CORRECT =  372           306
 INSERTIONS =    0             1
  DELETIONS =    1             0
  WORD RATE =      96.4%        97.1%
```

```
EXPERIMENT SUMMARY
speaker 45: female, 18, high school
          ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  25  50.0%     41  82.0%
  OPTION #2 =   5  10.0%      3   6.0%
  OPTION #3 =   0   0.0%      0   0.0%
  OPTION #4 =   0   0.0%      0   0.0%
  OPTION #5 =   0   0.0%      0   0.0%
     ERRORS =  20  40.0%      6  12.0%

 WORD TRIALS =  343          281
    CORRECT =  309           266
 INSERTIONS =    8             3
  DELETIONS =   13             2
  WORD RATE =      88.0%        93.7%
```

```
EXPERIMENT SUMMARY
speaker 46: female, 29, junior college
                 ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  34   68.0%      46   92.0%
  OPTION #2 =   3    6.0%       1    2.0%
  OPTION #3 =   5   10.0%       1    2.0%
  OPTION #4 =   1    2.0%       0    0.0%
  OPTION #5 =   2    4.0%       0    0.0%
     ERRORS =   5   10.0%       2    4.0%

 WORD TRIALS =  368            307
    CORRECT =  351             304
  INSERTIONS =    1              1
   DELETIONS =    2              0
   WORD RATE =       95.1%        98.7%
```

```
EXPERIMENT SUMMARY
speaker 47: female, 35, bachelors
                 ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  39   78.0%      46   92.0%
  OPTION #2 =   8   16.0%       3    6.0%
  OPTION #3 =   1    2.0%       0    0.0%
  OPTION #4 =   0    0.0%       0    0.0%
  OPTION #5 =   0    0.0%       0    0.0%
     ERRORS =   2    4.0%       1    2.0%

 WORD TRIALS =  368            307
    CORRECT =  356             302
  INSERTIONS =    1              0
   DELETIONS =    5              0
   WORD RATE =       96.5%        98.4%
```

```
EXPERIMENT SUMMARY
speaker 48: female, 36, junior college
                 ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
    CORRECT =  21   42.0%      25   50.0%
  OPTION #2 =   6   12.0%       3    6.0%
  OPTION #3 =   0    0.0%       1    2.0%
  OPTION #4 =   1    2.0%       0    0.0%
  OPTION #5 =   0    0.0%       0    0.0%
     ERRORS =  22   44.0%      21   42.0%

 WORD TRIALS =  369            306
    CORRECT =  313             262
  INSERTIONS =   10              7
   DELETIONS =    7              3
   WORD RATE =       82.6%        83.7%
```

```
EXPERIMENT SUMMARY
speaker 49: male, 24, junior college
              ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  32   64.0%     46   92.0%
   OPTION #2 =   2    4.0%      0    0.0%
   OPTION #3 =   1    2.0%      0    0.0%
   OPTION #4 =   1    2.0%      0    0.0%
   OPTION #5 =   1    2.0%      0    0.0%
      ERRORS =  13   26.0%      4    8.0%

  WORD TRIALS =  368            307
     CORRECT =  342             297
  INSERTIONS =   0               0
   DELETIONS =  16               2
   WORD RATE =       92.9%         96.7%
```

```
EXPERIMENT SUMMARY
speaker 50: male, 21, junior college
              ALL WORDS          SEMANTIC
PHRASE TRIALS =  50
     CORRECT =  24   48.0%     37   74.0%
   OPTION #2 =   4    8.0%      2    4.0%
   OPTION #3 =   1    2.0%      0    0.0%
   OPTION #4 =   0    0.0%      0    0.0%
   OPTION #5 =   0    0.0%      0    0.0%
      ERRORS =  21   42.0%     11   22.0%

  WORD TRIALS =  336            272
     CORRECT =  281             242
  INSERTIONS =  11               7
   DELETIONS =  13               0
   WORD RATE =       81.0%         86.7%
```

ATE

LMED

8